Sulla teoria generalizzata della gravitazione

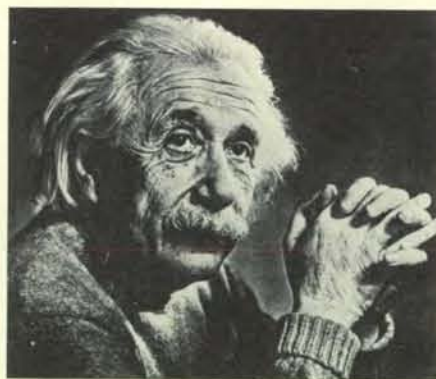
di Albert Einstein

Cade quest'anno il centenario della nascita di Albert Einstein. La stampa scientifica e non scientifica di tutto il mondo ha rievocato la figura, le opere e il contributo sostanziale che egli ha dato allo sviluppo della scienza attuale. Da parte nostra vogliamo ricordarlo pubblicando in italiano un articolo originale che Einstein scrisse per «Scientific American» e che apparve nel fascicolo di aprile del 1950.

La redazione di «Scientific American» mi ha chiesto di scrivere sul lavoro che ho pubblicato di recente. Si tratta di una ricerca matematica riguardante i fondamenti della fisica dei campi.

Qualche lettore potrebbe domandarsi se sui fondamenti della fisica non si sia già imparato tutto a scuola. La risposta può essere affermativa o negativa a seconda dell'interpretazione. Abbiamo appreso i concetti e le relazioni generali che ci mettono in grado di comprendere una immensa gamma di esperienze e di renderle accessibili al trattamento matematico. In un certo senso questi concetti e queste relazioni sono probabilmente anche definitivi. Questo è vero, per esempio, per le leggi della rifrazione della luce, per le relazioni della termodinamica classica per la parte basata sui concetti di pressione, volume, temperatura, calore e lavoro e per l'ipotesi della non esistenza di una macchina a moto perpetuo.

Che cosa ci spinge dunque a elaborare teoria dopo teoria? Perché, addirittura, formuliamo teorie? La risposta alla seconda domanda è semplice: perché amiamo «comprendere», ossia ridurre i fenomeni per mezzo del procedimento logico a qualcosa di già noto o (manifestante) evidente. Prima di tutto sono necessarie nuove teorie quando si affrontano fatti nuovi che non possono essere «spiegati» da teorie esistenti. Ma questa motivazione è, per così dire, banale, imposta dall'esterno. C'è un'altra motivazione più sottile e di non minore importanza. Si tratta dello sforzo verso l'unificazione e la semplificazione delle premesse della teoria nel suo insieme (ossia, il principio di economia di Mach, interpretato come un principio logico).

Esiste una passione per la comprensione proprio come esiste una passione per la

Albert Einstein (Camera Press).

musica; è una passione molto comune nei bambini, ma che poi la maggior parte degli adulti perde. Senza di essa non ci sarebbero né la matematica né altre scienze. Più volte la passione per la comprensione ha condotto all'illusione che l'uomo sia in grado di comprendere razionalmente il mondo oggettivo, attraverso il pensiero puro, senza nessuna fondazione empirica; in breve, attraverso la metafisica. Sono convinto che ogni teorico vero sia una sorta di metafisico addomesticato, indipendentemente da quanto possa immaginare di essere un puro «positivista». Il metafisico crede che il logicamente semplice sia anche reale. Il metafisico addomesticato crede che non tutto ciò che è logicamente semplice sia incorporato nella realtà esperita, ma che la totalità di tutta l'esperienza sensoriale possa essere «compresa» sulla base di un sistema concettuale costruito su premesse di grande semplicità. Lo scettico dirà che questo è un «credo del miracolo». È proprio così, ma è un credo del miracolo che è nato e cresciuto in maniera straordinaria grazie allo sviluppo della scienza.

La nascita dell'atomismo ne è un esempio efficace. Come può aver concepito Leucippo questa audace idea? Quando l'acqua gela e diventa ghiaccio - qualcosa all'apparenza interamente diverso dall'acqua - perché la fusione del ghiaccio dà origine a qualcosa che non sembra distinguersi dall'acqua originaria? Leucippo è perplesso e cerca una «spiegazione». Giunge alla conclusione che in queste transizioni l'«essenza» della cosa non è cambiata affatto. Può darsi che la cosa consista di particelle immutabili e che il cambiamento sia solo un cambiamento nella loro disposizione spaziale. Non potrebbe essere vero per tutti gli oggetti materiali che si manifestano ripetutamente con proprietà quasi identiche?

Questa idea non è andata del tutto perduta durante il lungo letargo del pensiero occidentale. Duemila anni dopo Leucippo, Bernoulli si chiede perché un gas esercita una pressione sulle pareti di un contenitore. Si potrebbe «spiegare» questo fenomeno con la mutua repulsione delle parti del gas, nel senso della meccanica newtoniana? Questa ipotesi sembra assurda, poiché la pressione del gas dipende dalla temperatura, a parità delle altre condizioni. Supporre che le forze newtoniane di interazione dipendano dalla temperatura è contrario allo spirito della meccanica newtoniana. Dato che a Bernoulli è nota la concezione atomista, è portato a concludere che gli atomi (o le molecole) collidono con le pareti del recipiente e così facendo esercitano una pressione. Dopo tutto, si deve supporre che gli atomi siano in movimento; in quale altro modo si potrebbe spiegare la variazione di temperatura dei gas?

Una semplice considerazione meccanica dimostra che questa pressione dipende soltanto dall'energia cinetica delle particelle e dalla loro densità nello spazio. Questo avrebbe dovuto far concludere ai fisici di quel tempo che il calore consiste nel moto casuale degli atomi. Se essi avessero tenuto questa considerazione nel debito conto, lo sviluppo della teoria del calore - in particolare la scoperta dell'equivalenza del calore e dell'energia meccanica - sarebbe stato notevolmente facilitato.

Questo esempio intende illustrare due cose. L'idea teorica (in questo caso l'atomismo) non nasce al di fuori e indipendentemente dall'esperienza; né può derivare dall'esperienza per puro procedimento logico. È il prodotto di un atto creativo. Una volta che l'idea teorica sia acquisita, è bene seguirla finché conduce a una conclusione insostenibile.

Per quanto riguarda il mio ultimo lavoro teorico, non mi sento giustificato a darne un resoconto dettagliato a un vasto gruppo di lettori interessati alla scienza. Ciò si dovrebbe fare solo con teorie che siano state adeguatamente confermate dall'esperienza. Fino a questo momento sono soprattutto la semplicità delle sue premesse e la sua intima connessione con ciò che già si conosce (e cioè le leggi del campo gravitazionale puro) che parlano a favore della teoria che mi accingo a esporre. Può darsi però che interessi un vasto gruppo di lettori conoscere il filo di pensiero che può portare a tentativi di natura così estremamente speculativa. Inoltre, si potrà vedere il tipo di difficoltà incontrate e come sono state superate.

Nella fisica newtoniana il concetto teorico elementare su cui si basa la descrizione teorica dei corpi materiali è il punto materiale, o particella. Quindi la materia è considerata discontinua *a priori*. Questo rende necessario considerare l'azione reciproca dei punti materiali come «azione a distanza». Poiché quest'ultimo concetto sembra del tutto contrario all'esperienza di ogni giorno, è naturale che i contemporanei di Newton - e Newton stesso - trovassero difficile accettarlo. A causa, però, del successo quasi miracoloso del sistema newtoniano, le generazioni successive di fisici si abituarono all'idea di azione a distanza. Ogni dubbio fu sepolto per molto tempo a venire.

Ma quando, nella seconda metà del XIX secolo, divennero note le leggi dell'elettrodinamica, si vide che queste leggi non potevano essere incorporate in maniera soddisfacente nel sistema newtoniano. È affascinante domandarsi: Faraday avrebbe scoperto la legge dell'induzione elettromagnetica se avesse ricevuto un'educazione scolastica normale? Sgombro dal modo tradizionale di pensare egli sentiva che l'introduzione del «campo» come un elemento indipendente della realtà lo aiutava a coordinare i fatti sperimentali. Fu Maxwell a comprendere pienamente il significato del concetto di campo; egli fece la scoperta fondamentale che le leggi dell'elettrodinamica trovavano la loro espressione naturale nelle equazioni differenziali per i campi elettrico e magnetico. Queste equazioni implicavano l'esistenza di onde, le cui proprietà corrispondevano a quelle della luce così come erano note a quel tempo.

Questo inserimento dell'ottica nella teoria dell'elettromagnetismo rappresenta uno dei maggiori trionfi nello sforzo verso l'unificazione dei fondamenti della fisica; Maxwell raggiunse questa unificazione per mezzo di argomentazioni puramente teoriche, molto prima che fosse convalidata dal lavoro sperimentale di Hertz. La nuova concezione consentì di fare a meno dell'ipotesi dell'azione a distanza, per lo meno nel dominio dei fenomeni elettromagnetici; il campo appariva ora come il solo portatore di interazione elettromagnetica tra corpi e il comportamento del campo era completamente determinato da processi contigui, espressi da equazioni differenziali.

A questo punto sorge una domanda: poiché il campo esiste perfino in un vuoto, si deve considerare il campo come lo stato di un «portatore», oppure gli si deve piuttosto attribuire un'esistenza indipendente, non riducibile a nessuna altra cosa? In altre parole, esiste un «etere» che porta il campo e che dovrebbe essere considerato in uno stato ondulatorio, per esempio, quando porta le onde luminose?

La domanda ha una risposta naturale. Poiché non si può fare a meno del concetto di campo, è preferibile non introdurre un altro portatore con proprietà ipotetiche. Tuttavia i pionieri che per primi riconobbero l'indispensabilità del concetto di campo erano ancora troppo imbevuti dalla mentalità meccanicistica per accettare senza esitazione questo semplice punto di vista. Ma nel corso dei decenni successivi questo punto di vista prese lentamente piede.

L'introduzione del campo come concetto elementare rese la teoria incoerente nel suo insieme. La teoria di Maxwell, anche se descrive in maniera soddisfacente il comportamento delle particelle elettricamente cariche nelle loro interazioni reciproche, non spiega il comportamento delle densità elettriche, cioè non fornisce una teoria delle particelle stesse, che devono quindi essere trattate come masse puntiformi sulla base della vecchia teoria. La combinazione dell'idea di un campo continuo con quella di punti materiali discontinui nello spazio appare incoerente. Una teoria dei campi coerente richiede la continuità di tutti gli elementi della teoria, non soltanto nel tempo ma anche nello spazio e in tutti i punti dello spazio. Di conseguenza la particella materiale non trova posto come concetto fondamentale nella teoria dei campi. Quindi, anche se non si tiene conto del fatto che la gravitazione non vi è inclusa, l'elettrodinamica di Maxwell non può essere considerata una teoria completa.

Le equazioni di Maxwell per lo spazio vuoto rimangono invariate se le coordinate spaziali e il tempo sono soggetti a un tipo particolare di trasformazioni lineari - le trasformazioni di Lorentz («covarianza» rispetto alle trasformazioni di Lorentz). La covarianza vale anche, naturalmente, per una trasformazione composta di due o più trasformazioni di questo tipo; ciò che viene detta la proprietà di «gruppo» delle trasformazioni di Lorentz.

Le equazioni di Maxwell implicano il «gruppo di Lorentz», ma il gruppo di Lorentz non implica le equazioni di Maxwell. Il gruppo di Lorentz in realtà potrebbe essere definito indipendentemente dalle equazioni di Maxwell come un gruppo di trasformazioni lineari che lasciano invariante un particolare valore della velocità: la velocità della luce. Queste trasformazioni valgono per la transizione da un «sistema inerziale» a un altro che è in moto uniforme relativamente al primo. La proprietà nuova di maggior rilievo di questo gruppo di trasformazioni è che distrugge il carattere assoluto del concetto di simultaneità di eventi distanti l'uno dall'altro nello spazio. Su questa base c'è da aspettarsi che tutte le equazioni della fisica siano covarianti rispetto alle trasformazioni di Lorentz (teoria speciale della relatività). Così ne risultava che le equazioni di Maxwell conducevano a un principio euristico valido molto al di là del campo di applicabilità o addirittura di validità delle equazioni stesse.

La relatività speciale (o ristretta) ha questo in comune con la meccanica newtoniana: si suppone che le leggi di entrambe le teorie siano valide soltanto rispetto a certi sistemi di coordinate, quelli noti come «sistemi inerziali». Un sistema inerziale è un sistema in uno stato di moto tale che i punti materiali «liberi da forza» entro di esso non sono accelerati rispetto al sistema di coordinate. Questa definizione però è vuota se non vi è un mezzo indipendente per riconoscere l'assenza di forze. Ma un tale mezzo non esiste se la gravitazione è considerata come un «campo».

Sia A un sistema uniformemente accelerato rispetto a un «sistema inerziale» I. I punti materiali, non accelerati rispetto a I, sono accelerati rispetto a A, con accelerazione uguale, per tutti i punti, in grandezza e direzione. I punti si comportano come se esistesse un campo gravitazionale rispetto ad A, dato che è una proprietà caratteristica del campo gravitazionale che l'accelerazione sia indipendente dalla natura particolare del corpo. Non vi è ragione alcuna per escludere la possibilità di interpretare questo comportamento come effetto di un «vero» campo gravitazionale *(principio di equivalenza)*. Questa interpretazione implica che A sia un «sistema inerziale», anche se è accelerato

rispetto a un altro sistema inerziale. (È essenziale per questo argomento che l'introduzione di campi gravitazionali indipendenti sia considerata giustificata anche se non sono definite masse che generano il campo. Per questa ragione, tale argomentazione non sarebbe stata convincente per Newton.) I concetti di sistema inerziale, la legge di inerzia e la legge di moto sono privati, quindi, del loro significato concreto - non solo nella meccanica classica ma anche nella relatività speciale. Inoltre, seguendo questa linea di pensiero, risulta che rispetto ad *A* il tempo non può essere misurato da orologi identici; anzi, in generale va perduto perfino il significato fisico immediato delle differenze di coordinate. Considerate tutte queste difficoltà, non si dovrebbe tentare, dopo tutto, di restare fermi al concetto di sistema inerziale, rinunciando al tentativo di spiegare il carattere fondamentale dei fenomeni gravitazionali che si manifestano nel sistema newtoniano come equivalenza tra massa inerte e massa gravitazionale? Coloro che credono nella comprensibilità della natura devono rispondere negativamente.

Questa è l'essenza del principio di equivalenza: per spiegare l'identità tra massa inerte e massa gravitazionale nell'ambito della teoria è necessario ammettere trasformazioni non lineari delle quattro coordinate. Ciò significa che il gruppo delle trasformazioni di Lorentz e quindi l'insieme dei sistemi di coordinate «permessi» deve essere ampliato.

Quale gruppo di trasformazioni di coordinate può allora essere sostituito al gruppo delle trasformazioni di Lorentz? I matematici suggeriscono una risposta basata sulle ricerche fondamentali di Gauss e di Riemann: vale a dire, che il sostituto appropriato è il gruppo di tutte le trasformazioni continue (analitiche) delle coordinate. Rispetto a queste trasformazioni la sola cosa che rimane invariante è il fatto che punti vicini hanno quasi le stesse coordinate; il sistema di coordinate esprime soltanto l'ordine topologico dei punti nello spazio (incluso il suo carattere quadridimensionale). Le equazioni che esprimono le leggi di natura devono essere covarianti rispetto a tutte le trasformazioni continue delle coordinate. Questo è il principio della relatività generale.

Il procedimento appena descritto supera una carenza dei fondamenti della meccanica che era stata già notata da Newton ed era stata criticata da Leibniz e, due secoli dopo, da Mach: l'inerzia si oppone all'accelerazione, ma accelerazione relativa a che cosa? Nell'ambito della meccanica classica, la sola risposta è: l'inerzia si oppone all'accelerazione *relativa allo spazio.* Si tratta di una proprietà fisica dello spazio: lo spazio agisce sugli oggetti, ma gli oggetti non agiscono sullo spazio. È probabilmente il significato più profondo dell'affermazione di Newton *spatium est absolutum.* Ma l'idea disturbava qualcuno, in particolare Leibniz, il quale non attribuiva una esistenza indipendente allo spazio ma lo considerava semplicemente una proprietà delle «cose» (contiguità di oggetti fisici). Se i giustificati dubbi di Leibniz avessero preso il sopravvento, a quel tempo, difficilmente questo sarebbe stato vantaggioso per la fisica, dato che i fondamenti empirici e teorici necessari per sfruttare la sua idea non erano disponibili nel XVII secolo.

Secondo la relatività generale, il concetto di spazio separato da ogni contenuto fisico non esiste. La realtà fisica dello spazio è rappresentata da un campo le cui componenti sono funzioni continue di quattro variabili indipendenti: le coordinate dello spazio e del tempo. È proprio questo particolare tipo di dipendenza che esprime il carattere spaziale della realtà fisica.

Poiché la teoria della relatività generale implica la rappresentazione della realtà fisica per mezzo di un campo *continuo*, il concetto di particelle o di punti materiali non può avere una parte fondamentale, né lo può il concetto di moto. La particella può soltanto apparire come una regione limitata dello spazio in cui la forza del campo o la densità di energia sono particolarmente alte.

Una teoria relativistica deve rispondere a due domande: 1) In che cosa consiste il carattere matematico del campo? 2) Quali equazioni sono valide per questo campo?

Per quanto riguarda la prima domanda: dal punto di vista matematico il campo è caratterizzato essenzialmente dal modo in cui le sue componenti si trasformano se è applicata una trasformazione di coordinate. Per quanto riguarda la seconda domanda: le equazioni devono determinare il campo *in misura sufficiente* soddisfacendo sempre ai postulati della relatività generale. Che questo requisito possa essere soddisfatto o meno, dipende dalla scelta del tipo di campo.

Il tentativo di comprendere le correlazioni tra i dati empirici sulla base di un tale programma così astratto può apparire a prima vista senza speranza. Il procedimento in realtà equivale a porre la domanda: qual è la proprietà più semplice che può essere richiesta, e quale è l'oggetto (campo) più semplice per cui tale proprietà si può richiedere conservando il principio della relatività generale? Considerato dal punto di vista della logica formale, il carattere duale della domanda appare disastroso, a parte poi la vaghezza del concetto di «semplice». Inoltre dal punto di vista della fisica non vi è niente che garantisca l'assunzione che una teoria che sia «logicamente semplice» debba essere anche «vera».

Eppure qualunque teoria è speculativa. Quando i concetti fondamentali di una teoria sono relativamente «vicini all'esperienza» (come, per esempio, i concetti di forza, pressione, massa), il suo carattere speculativo non è facilmente individuabile. Se, però, una teoria è tale da richiedere l'applicazione di procedimenti logici complessi per raggiungere conclusioni, a partire dalle premesse, che possano essere confrontate con l'osservazione, chiunque acquisisce la consapevolezza della natura speculativa della teoria. In tal caso sorge un sentimento quasi irresistibile di avversione in coloro che non hanno esperienza nell'analisi epistemologica e che non sono consapevoli della natura precaria del pensiero teorico in quei campi che sono loro familiari.

D'altro canto, si deve ammettere che una teoria ha un vantaggio considerevole se i suoi concetti base e le sue ipotesi fondamentali sono «vicini all'esperienza» ed è certamente giustificata una maggior fiducia in una teoria di questo tipo. Si corre meno il pericolo di andare completamente fuori strada, soprattutto perché ci vuole molto meno tempo e sforzo per invalidare tali teorie con l'esperienza. Eppure, via via che la profondità della nostra conoscenza aumenta, dobbiamo rinunciare a questo vantaggio nella nostra ricerca di semplicità logica e di uniformità nei fondamenti della teoria fisica. Si deve ammettere che la relatività è andata oltre le teorie fisiche precedenti nel rinunciare alla «vicinanza all'esperienza» dei concetti fondamentali allo scopo di raggiungere la semplicità logica. Questo vale già per la teoria della gravitazione ed è ancora più vero per quanto riguarda la nuova generalizzazione, che è un tentativo di comprendere le proprietà del campo totale. Nella teoria generale, il procedimento di derivare dalle premesse della teoria conclusioni che possano essere confrontate con dati empirici è così difficile che finora non si è ottenuto a questo proposito alcun risultato. A favore di questa teoria vi sono, a questo punto, la sua semplicità logica e la sua «rigidità». Rigidità significa in questo contesto che la teoria è o vera o falsa, ma non modificabile.

La maggior difficoltà interna che impedisce lo sviluppo della teoria della relatività è costituita dalla natura duale del problema indicata dalle due domande che abbiamo formulato. Questa dualità costituisce il motivo per cui lo sviluppo della teoria è avvenuto in due passi così distanziati fra loro nel tempo. Il primo di questi passi, la teoria della gravitazione, è basato sul principio di equivalenza discusso prima e si fonda sulla seguente considerazione: in base alla teoria della relatività speciale, la luce ha una velocità di propagazione costante. Se un raggio di luce nel vuoto parte da un punto, designato dalle coordinate x_1, x_2 e x_3 in un sistema di coordinate a tre dimensioni, al tempo x_4, esso viaggia come un'onda sferica e raggiunge un punto vicino (x_1+dx_1, x_2+dx_2, x_3+dx_3) al tempo x_4+dx_4. Introducendo la velocità della luce, *c*, otteniamo l'espressione:

$$\sqrt{dx_1^2 + dx_2^2 + dx_3^2} = c\,dx_4$$

che può essere scritta anche nella forma:

$$dx_1^2 + dx_2^2 + dx_3^2 - c^2dx_4^2 = 0$$

Questa espressione rappresenta una relazione oggettiva fra punti spazio-temporali vicini in quattro dimensioni, e vale per tutti i sistemi inerziali, una volta che le trasformazioni di coordinate siano ristrette a quelle della relatività speciale. La relazione perde la sua forma, tuttavia, se vengono ammesse trasformazioni continue arbitrarie delle coordinate, in accordo con il principio della relatività generale. La relazione, allora, assume la forma più generale:

$$\Sigma\, g_{ik}\, dx_i\, dx_k = 0.$$

Le g_{ik} sono certe funzioni delle coordinate che si trasformano in un particolare modo quando viene applicata una trasformazione continua delle coordinate. In base al principio di equivalenza, queste funzioni g_{ik} descrivono un tipo particolare di campo gravitazionale: un campo che può essere ottenuto per trasformazione dello spazio «libero da campo». Le g_{ik} soddisfano una particolare legge di trasformazione. Matematicamente parlando, esse sono le componenti di un «tensore» che gode di una proprietà di simmetria che si conserva in tutte le trasformazioni; questa proprietà simmetrica è espressa così:

$$g_{ik} = g_{ki}.$$

L'idea si presenta da sola: non possiamo attribuire un significato oggettivo a un tale tensore simmetrico, anche se il campo *non può* essere ottenuto dallo spazio vuoto della relatività speciale attraverso una pura trasformazione di coordinate? Anche se non posiamo aspettarci che descriva il campo di tipo più generale, un tale tensore simmetrico, può certamente descrivere il caso particolare del «campo gravitazionale puro». Così è evidente che genere di campo debba postulare la teoria generale della relatività, almeno per un caso speciale: un campo tensoriale simmetrico.

Pertanto rimane aperta solo la seconda domanda: che genere di legge di campo generale covariante deve essere postulata per un campo tensoriale simmetrico?

Questa domanda non era difficile, per la nostra epoca, poiché concetti matematici necessari erano già disponibili, sotto forma della teoria metrica delle superfici, creata un secolo addietro da Gauss ed estesa da Riemann a varietà con un numero arbitrario di dimensioni. Il risultato di questa ricerca puramente formale è stato sorprendente sotto molti punti di vista. Le equazioni differenziali che possono venire postulate come leggi di campo per le g_{ik} non possono essere di ordine inferiore al secondo, cioè devono contenere almeno la derivata seconda delle g_{ik} rispetto alle coordinate. Assumendo che nella legge di campo non appaiono derivate di ordine superiore al secondo, allora la legge *è determinata matematicamente dal principio di relatività generale.* Il sistema di equazioni può essere scritto nella forma:

$$R_{ik} = 0.$$

Le R_{ik} si trasformano nello stesso modo delle g_{ik}, cioè anch'esse formano un tensore simmetrico.

Queste equazioni differenziali sostituiscono completamente la teoria newtoniana del moto dei corpi celesti, purché le masse siano rappresentate come singolarità del campo. In altre parole, esse contengono la legge della forza così come la legge del moto, mentre eliminano i «sistemi inerziali».

Il fatto che le masse appaiono come singolarità indica che queste masse non possono essere spiegate a loro volta per mezzo di campi g_{ik} simmetrici, cioè «campi gravitazionali». Da questa teoria non si può dedurre neanche il fatto che esistano solo masse di gravità *positiva.* Evidentemente una teoria completamente relativistica dei campi deve basarsi su un campo di natura più complessa, cioè una generalizzazione del campo tensoriale simmetrico.

Prima di passare a considerare una tale generalizzazione, per seguire la spiegazione sono necessarie due osservazioni relative alla teoria della gravitazione.

La prima osservazione è che il principio di relatività generale impone alle possibilità teoriche delle restrizioni estremamente forti. Senza questo principio restrittivo, sarebbe praticamente impossibile per chiunque affrontare le equazioni gravitazionali, neppure usando il principio della relatività speciale, e anche se si sa che il campo deve essere descritto attraverso un tensore simmetrico. Nessuna raccolta di fatti, per estesa che sia, potrebbe condurre a queste equazioni, a meno che non si usi il principio della relatività generale. Questo è il motivo per cui tutti i tentativi di ottenere una conoscenza più profonda dei fondamenti della fisica mi sembrano destinati a fallire a meno che i concetti di base siano sin dall'inizio in accordo con la relatività generale. Questa situazione rende difficile utilizzare la nostra conoscenza empirica, per comprensiva che possa essere, nella ricerca dei concetti e delle realizzazioni fondamentali della fisica, e ci spinge ad applicare la libera riflessione in una misura molto superiore a quella attualmente assunta dalla maggior parte dei fisici. Non vedo alcun motivo per pensare che l'importanza euristica del principio di relatività generale sia ristretta alla gravitazione e che il resto della fisica possa essere affrontato separatamente sulla base della relatività speciale, con la speranza che possa poi essere inserito complessivamente, in modo coerente, in uno schema relativistico generale. Non penso che un tale atteggiamento, per quanto storicamente comprensibile, possa essere oggettivamente giustificato. La relativa piccolezza di quelli che oggi conosciamo come effetti gravitazionali non costituisce una ragione conclusiva per ignorare il principio della relatività generale in indagini teoriche di carattere fondamentale. In altre parole, non credo che sia giustificabile chiedere: che aspetto avrebbe la fisica senza la gravitazione?

Il secondo punto da notare è che le equazioni della gravitazione sono 10 equazioni differenziali per le 10 componenti del tensore simmetrico g_{ik}. Nel caso di una teoria relativistica non generale, ordinariamente un sistema non è sovradeterminato se il numero delle equazioni è uguale al numero delle funzioni incognite. La varietà delle soluzioni è tale che, entro la soluzione generale, un certo numero di funzioni a tre variabili può essere scelto arbitrariamente. Per una teoria relativistica generale, non ci si può aspettare che sia ovvio che le cose vadano così. La libera scelta rispetto al sistema di coordinate implica che delle dieci funzioni di una soluzione, cioè delle componenti del campo, quattro possano essere tali da assumere un determinato valore in seguito a una opportuna scelta del sistema di coordinate. In altre parole, il principio di relatività generale implica che il numero delle funzioni che devono essere determinate per mezzo di equazioni differenziali non è 10, ma 10 − 4 = 6. Per queste sei funzioni possono essere postulate solo sei equazioni differenziali indipendenti. Solo sei delle 10 equazioni differenziali del campo gravitazionale debbono essere indipendenti l'una dall'altra, mentre le quattro restanti debbono essere connesse a queste sei per mezzo di quattro relazioni (identità). E anzi esistono, fra i primi membri, R_{ik}, delle 10 equazioni gravitazionali, quattro identità - le «identità di Bianchi» - che assicurano la loro «compatibilità».

In un caso come questo - quando il numero delle variabili del campo è uguale al numero delle equazioni differenziali - la compatibilità è sempre assicurata se le equazioni possono essere ottenute da un principio di variazione. Questo è effettivamente il caso che ci si presenta con le equazioni gravitazionali.

Tuttavia, le 10 equazioni differenziali non possono essere sostituite integralmente da sei. Il sistema di equazioni è in effetti «sovradeterminato», ma, per l'esistenza delle identità è sovradeterminato in modo tale che la sua compatibilità non

va perduta, cioè la varietà delle soluzioni non è ristretta in maniera critica. Il fatto che le equazioni della gravitazione implichino la legge del moto per le masse è strettamente collegato con questa (ammissibile) sovradeterminazione.

Dopo questa preparazione è facile afferrare la natura dell'indagine attuale senza entrare nei suoi dettagli matematici. Il problema è quello di costruire una teoria relativistica per il campo totale. L'indizio più importante per la soluzione sta nel fatto che esiste già la soluzione per il caso speciale del campo gravitazionale puro. La teoria che stiamo cercando pertanto deve essere una generalizzazione del campo gravitazionale. La prima domanda è: qual è la generalizzazione naturale del campo tensoriale simmetrico?

Non si può rispondere a questa domanda da sola, ma unicamente in connessione con l'altra domanda: quale generalizzazione del campo fornirà il sistema teorico più naturale? La risposta su cui si basa la teoria in discussione è che il campo tensoriale simmetrico deve essere sostituito da un campo non simmetrico. Questo significa che deve cadere la condizione $g_{ik} = g_{ki}$ per le componenti del campo. In questo caso il campo ha 16 componenti indipendenti, anziché 10.

Rimane il compito di formulare le equazioni differenziali relativistiche per un campo tensoriale non simmetrico. Nel tentativo di risolvere questo problema si incontra una difficoltà che non sorge, invece, nel caso del campo simmetrico. Il principio di relatività generale non è sufficiente per determinare completamente le equazioni del campo, principalmente perché la legge di trasformazione della parte simmetrica del campo, da sola, non coinvolge le componenti della parte antisimmetrica, o viceversa. Questa è probabilmente la ragione per cui finora non si è quasi mai tentato di raggiungere questo tipo di generalizzazione del campo. Si può dimostrare che la combinazione delle due parti del campo è una procedura naturale solo se nel formalismo della teoria è solo il campo totale che svolge un ruolo preciso, e non le parti simmetrica e antisimmetrica separatamente.

Si è trovato che questo requisito può essere effettivamente soddisfatto in maniera naturale. Ma anche questo requisito, insieme con il principio di relatività generale, non è ancora sufficiente per determinare univocamente le equazioni di campo. Vorrei ricordare che il sistema di equazioni deve soddisfare un'ulteriore condizione: le equazioni devono essere compatibili. Si è detto prima che questa condizione è soddisfatta se le equazioni possono essere derivate da un principio di variazione.

In effetti si è riusciti a raggiungere questo risultato, anche se non in modo così naturale come nel caso del campo simmetrico. Ci ha lasciato perplessi il fatto di trovare che questo risultato può essere raggiunto in due modi differenti. Questi principi di variazione ci hanno dato due sistemi di equazioni - chiamiamoli E_1 ed E_2 - che differiscono l'uno dall'altro (sia pure di poco), e ciascuno dei quali presenta specifiche imperfezioni. Di conseguenza anche la condizione di compatibilità risultava insufficiente per determinare il sistema di equazioni in maniera univoca.

Di fatto, sono stati difetti formali dei sistemi E_1 ed E_2 che hanno indicato una possibile via d'uscita. Esiste un terzo sistema di equazioni, E_3, che è esente dai difetti formali dei sistemi E_1 e E_2 e che rappresenta una loro combinazione, nel senso che ogni soluzione di E_3 è una soluzione tanto di E_1 quanto di E_2. Questo fa pensare che E_3 sia il sistema che andavamo cercando. Perché non postulare E_3, allora, come sistema di equazioni? Un tale modo di procedere non è giustificato in assenza di un'ulteriore analisi, dal momento che la compatibilità di E_1 e quella di E_2 non implicano la compatibilità del sistema E_3, che è un sistema più forte, e in cui il numero delle equazioni supera di quattro il numero delle componenti del campo.

Una considerazione indipendente mostra che, senza tener conto del problema della compatibilità del sistema più forte, E_3 è l'unica generalizzazione davvero naturale delle equazioni della gravitazione.

Ma E_3 non è un sistema compatibile nello stesso senso in cui lo sono i sistemi E_1 e E_2, la cui compatibilità è assicurata da un numero sufficiente di identità, il che significa che ogni campo che soddisfi le equazioni per un ben definito valore del tempo possiede una estensione continua che rappresenta una soluzione nello spazio quadridimensionale. Il sistema E_3, tuttavia, non può essere esteso nello stesso modo. Usando il linguaggio della meccanica classica possiamo dire: nel caso del sistema E_3 le condizioni iniziali non possono essere scelte liberamente. Quel che importa davvero è la risposta alla domanda: la varietà delle soluzioni per il sistema E_3 possiede l'estensione che deve essere richiesta per una teoria fisica? Questo problema puramente matematico resta ancora irrisolto.

Lo scettico dirà: «Può ben essere vero che questo sistema di equazioni sia ragionevole da un punto di vista logico. Ma questo non prova che esso corrisponda alla natura.» Lei ha ragione, caro scettico. Solo l'esperienza può decidere della verità. Tuttavia abbiamo raggiunto qualcosa, se siamo riusciti a formulare una domanda dotata di significato e precisa. Dare una risposta affermativa o negativa non sarà facile, nonostante la grande quantità di fatti empirici noti. Derivare dalle equazioni conclusioni che possano essere messe a confronto con l'esperienza richiederà sforzi molto faticosi e probabilmente nuovi metodi matematici.

numero 129
maggio 1979
anno XII
volume XXII

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Gli oggetti Apollo

Conosciamo oggi una trentina di questi corpi, simili agli asteroidi, le cui orbite attraversano l'orbita terrestre: è possibile che si tratti di nuclei di comete che hanno perso i loro componenti volatili

di George W. Wetherill

Nel 1937 un oggetto del diametro di un chilometro circa, chiamato poi Hermes, passò a 800 000 chilometri dalla Terra, cioè al doppio circa della distanza Terra-Luna. Ancora non è stato rivisto. Pressappoco ogni secolo ci si può aspettare che un oggetto simile passi presso la Terra a una distanza minore di quella della Luna. Una volta ogni 250 000 anni in media un oggetto di quel genere cadrà sulla Terra. L'impatto libererà un'energia equivalente a 10 000 bombe all'idrogeno da 10 megaton e scaverà un cratere del diametro di 20 chilometri. Fortunatamente queste catastrofi sono così rare che non ne è stata registrata nessuna nel corso della storia umana.

Hermes, l'asteroide trovato nel 1937 mentre attraversava l'orbita terrestre, era solo il terzo corpo di quel genere scoperto. Il primo di quegli oggetti era stato scoperto cinque anni prima, nel 1932, da Karl Reinmuth dell'Università di Heidelberg nel corso di una ricerca fotografica di asteroidi comuni. Oggetti analoghi saranno stati fotografati certamente anche in precedenza, ma non erano stati identificati come tali. Il moto veloce di un corpo che si avvicina alla Terra lascia una traccia sulla lastra fotografica che non è distinguibile a prima vista dalla traccia di una meteora o da una imperfezione della lastra. Reinmuth fu il primo a dimostrare che l'oggetto da lui stesso fotografato era un asteroide, la cui orbita ellittica attraversava l'orbita terrestre prima di raggiungere il suo perielio (la minima distanza dal Sole). L'oggetto fu indicato provvisoriamente come 1932 HA e in seguito fu chiamato Apollo, dal nome della divinità solare della mitologia greca. Più tardi tutti i corpi le cui orbite attraversano l'orbita terrestre furono chiamati genericamente «oggetti Apollo».

Per caso, nello stesso anno fu scoperto un altro piccolo asteroide, Amor, con un'orbita simile a quella di Apollo, ma col perielio appena oltre l'orbita terrestre, a 1,08 unità astronomiche dal Sole. (Un'unità astronomica, abbreviata UA, è definita come la distanza media tra la Terra e il Sole.) In seguito si scoprì che le orbite degli oggetti con perielio così vicino a 1 UA si evolvono spesso in orbite che tagliano quella terrestre. Ora gli astronomi chiamano «oggetti Apollo-Amor» tutti i corpi con un perielio minore del valore fissato arbitrariamente di 1,3 UA.

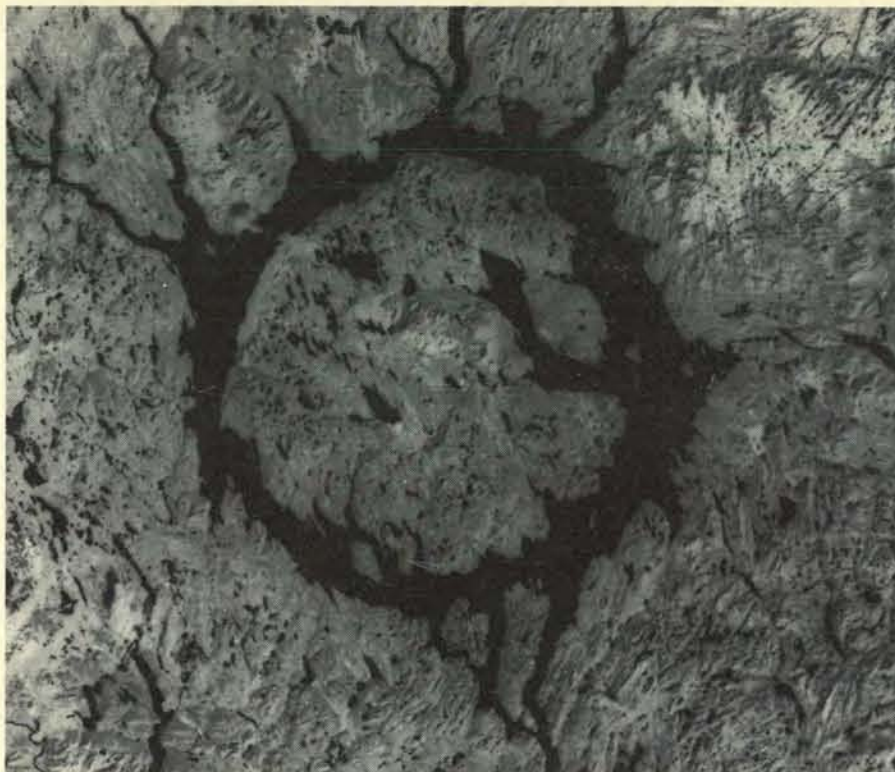
Dal 1932 sono stati identificati circa 28 oggetti Apollo le cui orbite tagliano quella terrestre e un numero un po' inferiore di oggetti Amor. Si deve dire «circa», perché in qualche caso l'orbita non è stata determinata con precisione tale da garantirne l'appartenenza alla classe Apollo-Amor. Anche quando sembrò che l'orbita fosse ben definita fu destino comune dei primi oggetti Apollo di essere presto perduti. Lo stesso Apollo non fu ritrovato fino al 1973, e Adonis, che fu il secondo identificato (nel 1936) di questi corpi, non fu rivisto per 40 anni, durante i quali attraversò l'orbita terrestre più di 30 volte.

Fino a poco tempo fa si prestò poca attenzione agli oggetti Apollo-Amor, anche da parte della minoranza di astronomi specializzati negli studi planetari. Solo lentamente si capì che questi piccoli asteroidi sono molto più importanti per la conoscenza terrestre e planetaria di quanto lascino supporre le loro dimensioni e il loro numero. I corpi di tipo Apollo-Amor sono stati i principali responsabili dei crateri di diametro superiore a cinque chilometri sulla Terra, la Luna, Mercurio, Venere e Marte (con la possibile eccezione di Marte). A sua volta la densità superficiale di questi crateri costituisce il mezzo principale per la determinazione dell'età relativa di regioni diverse di questi vicini spaziali.

Una notevole percentuale, probabilmente la maggior parte, delle meteoriti che cadono sulla Terra è costituita da frammenti di oggetti Apollo prodotti nelle collisioni che hanno luogo quando questi oggetti, seguendo le loro orbite eccentriche, finiscono nella fascia degli asteroidi, che si trova tra le orbite di Marte e di Giove. Le meteoriti sono le fonti più importanti di dati dettagliati sulla storia primordiale del sistema solare. L'interpretazione di questi dati richiede informazioni sulle sorgenti delle meteoriti.

Si può affermare con una certa sicurezza che gli oggetti Apollo-Amor non sono semplicemente un campione statistico di oggetti espulsi in qualche modo dall'ampia fascia degli asteroidi, che occupa una regione compresa pressappoco tra 2,1 e 3,5 UA. Il primo asteroide, Cerere, fu scoperto nel 1801 dal monaco italiano Giuseppe Piazzi. Cerere, che è il più grande asteroide conosciuto, ha un diametro di 1000 chilometri circa e si muove in un'orbita a circa 2,8 UA dal Sole. Finora sono stati catalogati oltre 2000 asteroidi, la grande maggioranza dei quali si trova oltre l'orbita di Marte (1,5 UA). Si pensa che la fascia degli asteroidi comprenda circa 400 000 oggetti di diametro superiore a un chilometro. Come vedre-

Apollo, il primo oggetto Apollo conosciuto, fu scoperto nel 1932. Poi fu perduto. Queste stampe in negativo ne registrano la riscoperta dopo 41 anni, il 28 marzo 1973. Apollo è la debole immagine arrotondata al centro nella fotografia di sinistra. Nella fotografia di destra, scattata un'ora dopo, si è mosso verso nord-ovest. Le stelle luminose circostanti, che appartengono alla costellazione della Vergine, appaiono come strisce, dato che nei 25 minuti di esposizione il telescopio era guidato alla velocità prevista per l'oggetto Apollo. Se si fossero tenute fisse le stelle, il debole oggetto Apollo, che si trovava allora a 100 milioni di chilometri dalla Terra, non avrebbe lasciato nessuna traccia visibile. Le fotografie del ritrovamento furono ottenute col riflettore da 155 centimetri del Harvard College Observatory da Richard F. McCrosky e Cheng-Yuan Shao, che si erano preparati per una ricerca di vari mesi. Per una fortuita coincidenza Apollo fu trovato nella prima fotografia.

2100 Ra-Shalom, che lasciò la breve traccia su questa stampa in negativo, fu il terzo dei quattro oggetti Apollo scoperti l'anno scorso. Fu trovato da Eleanor F. Helin del California Institute of Technology su una lastra impressionata il 10 settembre col telescopio Schmidt di Monte Palomar. La scoperta fu in realtà una riscoperta di un oggetto già fotografato vari anni prima e in seguito perduto. Il numero 2100 indica che questo oggetto è il 2100-esimo elencato nel catalogo ufficiale di tutti gli asteroidi di qualsiasi specie. I nomi «Ra», nome dell'antica divinità solare egizia, e «Shalom», parola ebraica che significa pace, furono attribuiti all'oggetto al tempo dell'incontro di Camp David.

Cratere gigante nel Quebec nordorientale formato circa 210 milioni di anni fa dall'impatto di un corpo di grandi dimensioni simile agli oggetti Apollo. Per scavare un cratere come questo, che si estende per circa 70 chilometri, l'asteroide deve avere avuto un diametro di tre chilometri circa. La fotografia fu scattata dall'Earth Resources Technology Satellite (ERTS). La visibilità del cratere è accresciuta dall'acqua scura del lago Manicouagan, che ne riempie oggi la depressione circolare. Al centro della formazione c'è un picco di rocce che hanno subito «shock». Molti crateri nel Canada furono conservati sotto strati di sedimenti, poi trascinati via dall'azione glaciale.

mo, alcuni di questi oggetti possono diventare di tipo Apollo. D'altra parte, un'ipotesi recente e forse sorprendente prevede che molti oggetti Apollo-Amor non siano in realtà asteroidi, ma i resti privi di gas di comete, gli oggetti più primitivi del sistema solare.

Nell'ultimo decennio l'interesse per gli oggetti Apollo-Amor si è molto accresciuto. Quando gli astronomi osservano sulle loro lastre fotografiche un oggetto in rapido movimento, si affrettano spesso a comunicare la scoperta all'Ufficio Centrale per i Telegrammi Astronomici dell'Unione Astronomica Internazionale e ad altri astronomi. Si calcolano e si diffondono le effemeridi che indicano la posizione, la velocità e la direzione di movimento, in modo da consentire successive osservazioni. Talvolta oggetti appena scoperti possono essere riconosciuti su vecchie lastre. L'identificazione di un oggetto su varie lastre consente solitamente di calcolare la sua orbita con precisione sufficiente da assicurarne il ritrovamento in occasione dei suoi avvicinamenti successivi.

Il primo programma osservazionale per una ricerca sistematica di oggetti Apollo--Amor fu intrapreso nel 1973 all'Osservatorio di Monte Palomar da Eugene M. Shoemaker e Eleanor F. Helin del California Institute of Technology. Lavorando circa quattro notti al mese con un telescopio Schmidt a grande campo di dimensioni medie hanno scoperto cinque oggetti Apollo e un oggetto Amor, e altri corpi interessanti, tra cui una nuova cometa.

Contributi significativi sono stati portati anche da vari altri programmi di osservazione che utilizzavano telescopi Schmidt di maggiori dimensioni, ma che non erano rivolti espressamente alla ricerca di oggetti Apollo-Amor. Charles Kowal ha scoperto cinque oggetti Apollo-Amor lavorando col telescopio Schmidt da 1,20 metri di Monte Palomar, e altri tre sono stati scoperti da un gruppo di ricercatori che sta effettuando rilevamenti su tutto il cielo all'Osservatorio Europeo Australe in Cile.

A differenza di una decina di anni fa, oggi si sa che quasi metà degli oggetti Apollo-Amor sono certamente asteroidi ed è stato attribuito loro un numero. Pochi sono perduti senza speranza. La maggior parte della quarantina di oggetti presenta orbite molto eccentriche che si spingono nella fascia principale degli asteroidi e, in una mezza dozzina di casi, anche oltre. Almeno tre oggetti, però, hanno orbite moderatamente eccentriche, con un afelio (massima distanza dal Sole) interno all'orbita di Marte.

Qualsiasi discussione quantitativa sull'importanza di corpi Apollo-Amor per la formazione di crateri sui pianeti o per la produzione di meteoriti richiede una conoscenza della loro abbondanza. Il loro diametro angolare è nettamente inferiore alla migliore risoluzione angolare dei telescopi, pertanto le loro dimensioni devono essere determinate indirettamente. Questo è possibile misurando la loro albedo geometrica: cioè, la frazione di luce solare incidente che riflettono. Combinando l'albedo di un oggetto e la sua luminosità misurata con la sua distanza dalla Terra e dal Sole si può calcolare l'area della sua sezione, e da questa, supponendo che abbia una forma grossolanamente sferica, si può risalire alle sue dimensioni approssimate.

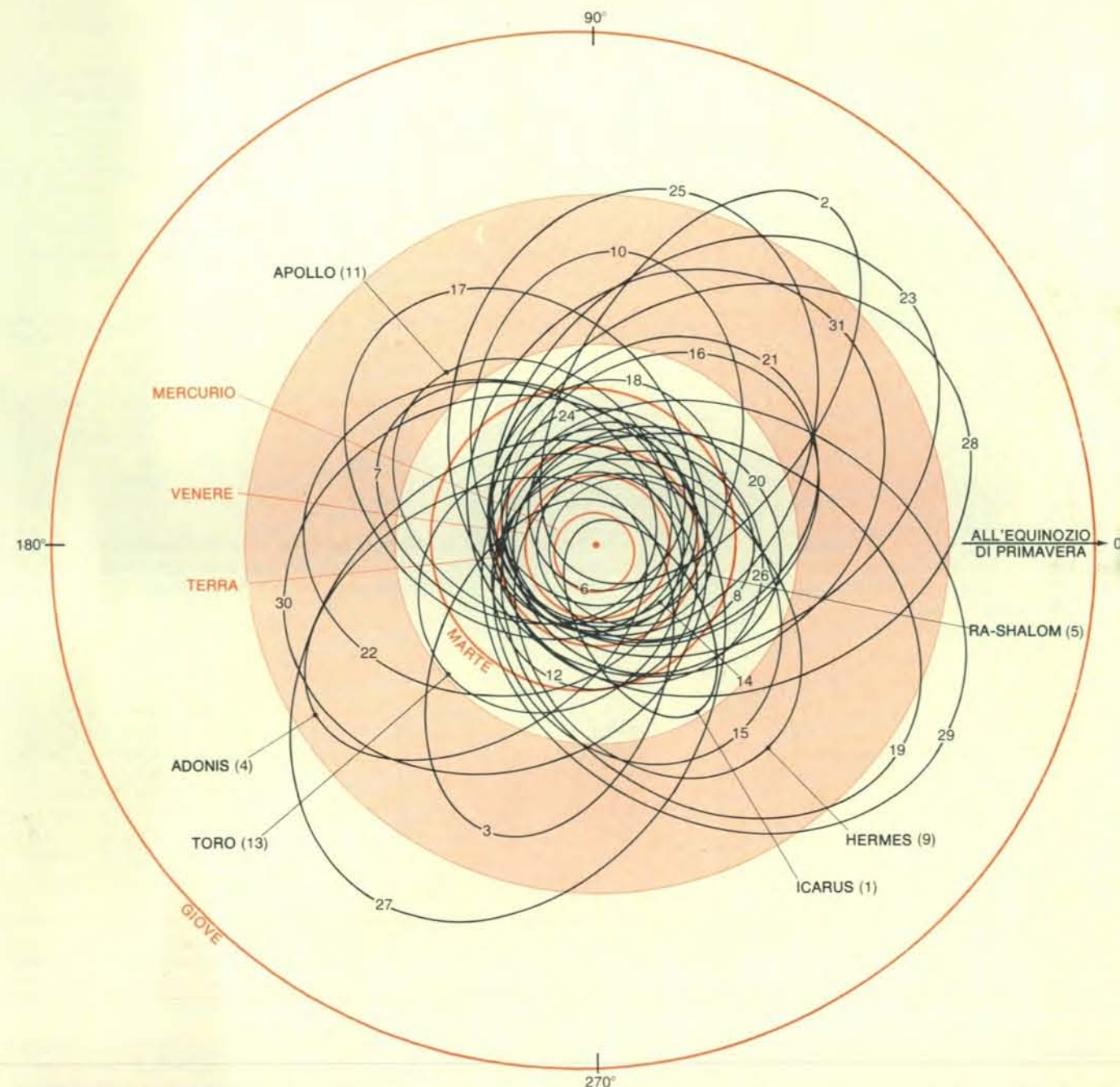
L'albedo può essere determinata in molti modi. In una situazione stazionaria l'energia incidente su un oggetto che non viene riflessa deve essere assorbita e poi riirradiata, per lo più nell'infrarosso. Pertanto, misurando la luminosità di un oggetto nell'infrarosso e confrontandola con la luminosità nel visibile, si può determinare la riflettività dell'oggetto. Misure di laboratorio hanno dimostrato anche che il grado di polarizzazione della luce riflessa da una superficie dipende sia dall'angolo di riflessione che dall'albedo. Inoltre il «colore» di un asteroide può essere determinato misurandone la luminosità in corrispondenza a lunghezze d'onda differenti: nella regione ultravioletta dello spettro, in quella visibile e in quella dell'infrarosso vicino. Spesso queste misure consentono di riconoscere almeno indicativamente negli asteroidi gruppi di minerali trovati nelle meteoriti e nelle rocce terrestri e lunari. Misure di laboratorio sull'albedo di questi gruppi di minerali possono darci un controllo ulteriore sull'albedo di un asteroide.

È facile trovare difetti in questi metodi per la determinazione dell'albedo degli asteroidi, ma va notato che in molti casi l'albedo è stata determinata facendo uso di più metodi e che quasi tutti i risultati sono in buon accordo tra loro. La maggior parte delle misure è stata effettuata su oggetti tipici della fascia degli asteroidi. Finora è stata misurata l'albedo solo di 11 oggetti Apollo e Amor. Comunque, le albedo geometriche di questo piccolo campione sono simili a quelle dei normali asteroidi e variano da 0,05 a 0,2. (Per confronto, l'albedo geometrica della Luna è 0,12.) Se si assume 0,16 come valore tipico dell'albedo geometrica degli oggetti Apollo-Amor, si trova che gli oggetti finora scoperti hanno diametri compresi per lo più tra uno e due chilometri. Un oggetto del diametro di un chilometro con un'albedo 0,16 posto a 1 UA di distanza sia dal Sole che dalla Terra avrà una magnitudine astronomica di + 18. Si dice magnitudine assoluta di un asteroide la sua magnitudine riferita a queste condizioni standard.

Si può sistemare il numero totale di oggetti Apollo-Amor più brillanti della magnitudine assoluta + 18 (cioè, con un diametro superiore a un chilometro) confrontando il numero di quelli effettivamente scoperti con una valutazione del grado di completezza raggiunto dalla ricerca telescopica. Il primo tentativo in questa direzione fu compiuto da Ernst J.

Öpik dell'Armagh Observatory in Irlanda nel 1963, quando erano noti 10 oggetti Apollo. Öpik ne trasse la conclusione che dovevano esserci almeno 43 oggetti Apollo, ma che potevano essercene molti di più. Poiché sono stati scoperti 28 oggetti Apollo e la frequenza delle nuove scoperte, ora di quattro circa all'anno, va aumentando, devono essercene effettivamente molti più di 43.

Una stima più recente sulla base del metodo della completezza di ricerca è stata effettuata da Shoemaker, che ha concluso che il numero totale di oggetti Apollo più brillanti della magnitudine + 18 è di 750 ± 300. Nel 1967 Fred L. Whipple dell'Università di Harvard considerò un metodo di stima piuttosto differente, basato sull'ipotesi che se il numero totale degli oggetti Apollo è piccolo, allora sarebbe molto probabile la riscoperta casuale di alcuni di essi. Fino al 1967 non si erano avute riscoperte casuali. Dall'algebra delle combinazioni e delle permutazioni si può dedurre un limite inferiore per il numero totale di oggetti Apollo. Dapprima Whipple stimò che fossero più di 50, e una ripetizione del calcolo nel 1973 diede un valore di 100. Sulla base dei dati più recenti, che comprendono la riscoperta casuale di un oggetto Apollo (Ra-Shalom) avvenuta nel 1978, il metodo rifiuta decisamente valori inferiori a 200 e porta a una stima in accordo con quella di Shoemaker.

Il fatto che un oggetto Apollo colpirà o no la Terra dipende dal caso. Al momento nessuno degli oggetti Apollo conosciuti è in rotta di collisione con la Terra. Tutti gli oggetti Apollo-Amor sono però soggetti continuamente all'attrazione

Le orbite di 31 oggetti Apollo sono rappresentate per confronto con le orbite di Mercurio, Venere, Terra, Marte e Giove. Sono numerate progressivamente al crescere della distanza dal Sole al perielio, cioè al momento del massimo avvicinamento al Sole. In questo stesso ordine sono elencati nella tabella a pagina 64. L'asteroide che passa più vicino al Sole è Icarus. Solo tre oggetti Apollo (Ra-Shalom, il numero 6 e il 14) hanno orbite contenute completamente entro l'orbita di Marte. La maggior parte segue orbite che attraversano la fascia principale degli asteroidi *(fascia larga in colore)*, in cui si pensa che si trovi circa mezzo milione di corpi di diametro superiore al chilometro. Frammenti di oggetti Apollo che hanno subito urti in tale fascia possono costituire un'ampia frazione delle meteoriti che cadono sulla superficie terrestre.

gravitazionale dei pianeti circostanti, in particolare di Giove, che determina la precessione delle orbite degli asteroidi. A seguito della precessione l'asse maggiore di un'orbita ellittica ruota progressivamente di 360 gradi nello spazio. Questo significa che tutti gli oggetti con un perielio interno all'orbita terrestre e un afelio esterno a essa finiranno per intersecarla prima o poi. Pertanto, un oggetto Apollo tipico si troverà ogni 5000 anni circa in un'orbita che taglia l'orbita terrestre.

Nella maggior parte dei casi la Terra sarà in qualche altro punto della sua orbita nel momento in cui l'oggetto Apollo passerà per il punto di intersezione, così che i due corpi non entreranno in collisione. Esiste però una probabilità calcolabile che la Terra e l'oggetto Apollo giungano contemporaneamente al punto di intersezione. La probabilità di collisione per un dato oggetto Apollo è pari soltanto a circa 5×10^{-9} per anno, equivalente cioè a una collisione in 200 milioni di anni. Se supponiamo che ci siano da 750 a 1000 oggetti Apollo con un diametro maggiore di un chilometro, deduciamo che circa quattro di essi colpiranno la Terra ogni milione di anni. Questa frequenza di collisione è piccola se confrontata coi 5000 anni della storia scritta dell'uomo, ma è notevole se confrontata coi miliardi di anni del tempo geologico.

Nel corso della storia scritta dell'uomo non c'è stata nessuna catastrofe prodotta dall'impatto di un oggetto Apollo. L'incontro più recente che si conosca con un oggetto di tipo asteroidale di dimensioni significative ebbe luogo tra 25 000 e 50 000 anni fa, quando un corpo ferroso di tipo asteroidale, che probabilmente non misurava più di 100 metri di diametro, scavò un cratere di poco più di un chilometro di diametro e profondo 200 metri: Il Meteor Crater in Arizona. Tra i crateri conosciuti che hanno un diametro superiore ai 10 chilometri, il più recente fa da bacino per il lago Bosumtwi nel Ghana. L'evento che lo generò avvenne circa 1,3 milioni di anni fa. I crateri piccoli come il Meteor Crater sono erosi rapidamente. Solo i crateri più grandi lasciano qualche traccia dopo 500 000 anni. Senza gli effetti di erosione prodotti dagli agenti atmosferici, dai processi geologici e dagli organismi viventi, la Terra apparirebbe coperta di crateri come la Luna.

Quando queste collisioni rare, ma statisticamente inevitabili, avvengono, si hanno conseguenze impressionanti. Nel sistema di riferimento della Terra l'energia cinetica di un oggetto Apollo di un chilometro di diametro e di densità normale (cioè, 3,5 grammi per centimetro cubo), che si muove in un orbita pure normale è di circa 4×10^{27} erg. Questa energia è pari a 100 000 volte l'energia liberata nell'eplosione di una testata nucleare da un megaton. Tutta questa energia deve essere dissipata in qualche modo nell'impatto con la Terra. Una piccola frazione dell'energia totale è sufficiente per vaporizzare il corpo proiettile, così che la maggior parte dell'energia sarà spesa per polverizzare il materiale della superficie terrestre e scagliarlo via ad alta velocità. Pertanto l'impatto scaverà un cratere simile a quelli che dominano i paesaggi sulla Luna, su Marte e su Mercurio.

Non ci sono dati diretti sulle dimensioni di crateri prodotti in esplosioni da 10^5 megaton. Si può però estrapolare qualche legge empirica basata su esplosioni molto più ridotte e simulare al calcolatore esplosioni di scala asteroidale. Secondo tali calcoli un impatto capace di liberare un'energia di 4×10^{27} erg scaverà un cratere del diametro di 22 chilometri circa.

Metà circa dei crateri antichi sicuri sono stati trovati nel Canada centrale e orientale, cioè in una regione che rappresenta solo l'1 per cento di tutte le terre emerse. I cerchi pieni indicano la posizione e le relative dimensioni di 23 crateri da impatto sicuri, numerati progressivamente in ordine di scoperta; i circoletti vuoti indicano possibili strutture da impatto. Il cratere più giovane conosciuto (1) ha meno di cinque milioni di anni; il più vecchio (13, presso Sudbury, Ontario) ha circa 1,8 miliardi di anni. Quest'ultimo è uno dei due soli crateri conosciuti di cui ci resta traccia dall'era Precambriana, che terminò 600 milioni di anni fa. Il cratere 3 è la formazione di Manicouagan, rappresentata nella fotografia aerea di pagina 14. Richard Grieve del Dipartimento canadese per l'energia , le miniere e le risorse ha fatto un conto complessivo degli antichi crateri sicuri nell'America Settentrionale e in Europa; in base a questo conto si può stimare che negli ultimi 600 milioni di anni caddero sulla Terra 1500 oggetti Apollo di diametro pari almeno a un chilometro, il 70 per cento dei quali cadde probabilmente negli oceani. La mappa si basa su una compilazione fatta dalla Divisione di fisica terrestre del Dipartimento per l'energia, le miniere e le risorse.

Se si accetta la stima in base alla quale la Terra sarebbe colpita circa quattro volte ogni milione di anni da oggetti capaci di produrre crateri di venti chilometri di diametro, ci si può chiedere dove se ne possano trovare le prove sperimentali. Ci si potrebbe attendere che la Terra sia abbondantemente ricoperta di crateri. È proprio così, ma fino a poco tempo fa tutto questo era sfuggito all'attenzione.

Un'attenta osservazione di fotografie aeree rivela molte strutture circolari, spesso poco evidenti, nel paesaggio di tutto il mondo. Studi geologici dettagliati hanno mostrato che si tratta in parte di resti di crateri da impatto, originariamente di diametro maggiore di 20 chilometri. La loro origine è provata dalla scoperta di strutture geologiche che ci si aspetta di trovare in crateri da esplosione fortemente erosi. Ci sono anche chiare prove che le rocce nella regione circostante hanno subito «shock» fortissimi. I dati di fatto comprendono la presenza di rocce ignee che si sono ricristallizzate dopo la fusione prodotta dall'impatto improvviso, e la presenza di tipi di quarzo, come la coesite e la stisciovite, che si generano solo a seguito di una combinazione di temperatura e pressione elevate. Tra le prove sperimentali macroscopiche ci sono i «coni di frantumazione», strutture di quarzite che si diramano verso l'esterno e verso il basso dalla direzione di impatto. In qualche caso le caratteristiche chimiche di un corpo non terrestre caduto sono riconoscibili nel materiale espulso dal cratere. Solo raramente, però, si trovano frammenti proprio dell'oggetto caduto, dato che è stato vaporizzato per la temperatura raggiunta nell'esplosione.

Questi indizi di origine per urto hanno portato oggi all'identificazione di oltre 50 strutture da impatto per lo più, non sorprendentemente, in regioni della Terra abitate da geologi interessati a esse. Metà circa dei crateri sicuri sono stati trovati nel Canada centrale e sudorientale, che costituisce soltanto l'1 per cento circa della superficie terrestre. Dato che non ci sono modi naturali per dirigere corpi extraterrestri verso una piccola regione della Terra, l'abbondanza di crateri nel Canada può essere attribuita soltanto all'efficacia del programma di ricerca predisposto dal Dipartimento canadese per l'energia, le miniere e le risorse di Ottawa.

Coni di frantumazione, cioè rocce che sono state frantumate secondo uno schema conico a seguito di un impatto, sono state trovate nella maggior parte delle sedi sicure di crateri. I coni visibili in questa fotografia sono stati trovati in un affioramento sulla riva meridionale del lago Keeley, presso il bacino di Sudbury: il resto di un cratere che si pensa abbia avuto un diametro originale di 140 chilometri. La struttura è quella indicata con il numero 13 nella mappa della pagina a fronte.

La maggior parte dei crateri ben studiati è costituita da crateri geologicamente giovani, cioè di età inferiore a poche centinaia di milioni di anni. Si conoscono solo due crateri precambriani (uno presso Sudbury in Ontario e l'altro presso Vredefort in Sud Africa), nonostante che il periodo precambriano comprenda più dell'80 per cento della storia terrestre. A meno che un cratere sia insolitamente grande, tutte le tracce della sua presenza saranno cancellate dall'erosione in meno di 600 milioni di anni. Inoltre, i crateri prodotti da asteroidi caduti negli oceani, che coprono più del 70 per cento della superficie terrestre, sono comprensibilmente di difficile identificazione. Tenendo conto di tutto si riconosce oggi che il tasso di formazione di crateri per unità di area sulla Terra è elevato almeno quanto quello sulla Luna.

Molti crateri terrestri identificati non superano i 100 metri di diametro e perciò cadono nell'intervallo di dimensioni proprie dei crateri che possono essere scavati dalle meteoriti più grandi conosciute. Possono essere chiamati correttamente crateri da meteoriti. Metà circa dei crateri terrestri conosciuti hanno diametri superiori a 10 chilometri, e pertanto devono essere stati scavati da corpi molto più grandi di qualsiasi meteorite trovata finora. In altre parole, questi crateri devono essere stati prodotti dall'impatto di corpi di dimensioni comparibili a quelle degli oggetti Apollo visibili telescopicamente. Una stima plausibile basata sulla distribuzione dei crateri conosciuti prevede che circa 1500 oggetti di tipo Apollo abbiano colpito la Terra dopo la fine dell'era precambriana 600 milioni di anni fa. Tenendo conto dell'incertezza attuale sul numero totale, sulle dimensioni e sulla velocità degli oggetti Apollo, questa valutazione concorda ragionevolmente con la stima indicata in precedenza di quattro collisioni ogni milione di anni.

Come si può essere sicuri che è corretto correlare i grandi crateri terrestri e lunari con gli oggetti Apollo conosciuti solo recentemente? La risposta sta nel fatto che oggetti capaci di scavare crateri così grandi devono essere abbastanza grossi da risultare visibili coi telescopi. Ci sono solo due tipi di oggetti astronomici osservabili che attraversano l'orbita terrestre: gli oggetti Apollo e le comete. A differenza degli asteroidi, il nucleo solido di una cometa contiene ghiaccio e altre sostanze che volatilizzano quando la cometa passa vicino al Sole e producono una chioma nebulare, o testa, e spesso anche una coda. Le chiome si espandono frequentemente fino a raggiungere un diametro di migliaia di chilometri. Si può stimare il diametro del nucleo nella chioma confrontando i valori osservati e calcolati della quantità di gas che ci si aspetta che venga emessa, in funzione della distanza della cometa dal Sole. Un indizio più diretto per valutare le dimensioni del nucleo di una cometa è fornito dalla luminosità, quando dista così tanto dal Sole che manca la chioma, la quale altrimenti ne provoca un oscuramento. Queste stime indicano che i nuclei delle comete

hanno dimensioni sostanzialmente analoghe a quelle dedotte per gli oggetti Apollo: da uno a 10 chilometri.

Dato che sia gli oggetti Apollo sia le comete possono entrare in collisione con la Terra e la Luna lasciando dei crateri, ci si chiede come si possa valutare la frequenza di bombardamento per i due tipi di oggetti. Se si considerano solo le stime delle dimensioni e del numero degli oggetti Apollo, sembra che questi bastino per spiegare il numero di grandi crateri prodotti sulla Luna, dopo che i magmi basaltici si sono riversati nelle regioni dei «mari» lunari circa tre miliardi di anni fa. Come abbiamo visto, l'abbondanza stimata degli oggetti Apollo concorda anche col numero di grandi crateri che si ritiene siano stati prodotti sulla Terra a partire dal Precambriano.

Inoltre, si può dimostrare che se i nuclei delle comete non sono più grandi degli oggetti Apollo, le comete possono spiegare solo una piccola percentuale dei crateri di grandi dimensioni osservati sulla Luna. La percentuale sarebbe però molto maggiore se le comete avessero in realtà dimensioni più grandi di quelle stimate comunemente, e se, a seguito di ciò, ci fossero molte comete non osservate di diametro superiore a un chilometro. In tal caso gli impatti di comete potrebbero spiegare fino al 35 per cento dei crateri lunari (e terrestri) più grandi. Recentemente A. H. Delsemme dell'Università di Toledo e L'ubor Kresák dell'Istituto per l'astronomia di Bratislava in Cecoslovacchia hanno presentato indipendentemente prove del fatto che il numero di comete deboli sarebbe in realtà minore di quanto ci si potrebbe attendere. Questo significherebbe che il contributo delle comete alla formazione dei crateri è effettivamente basso.

Dopo aver operato questa distinzione tra gli oggetti Apollo e le comete, ora la renderò un po' confusa indicando una possibilità sulla quale ritornerò più dettagliatamente. Come fu suggerito inizialmente da Öpik, molti oggetti Apollo potrebbero essere comete «estinte», cioè i resti non volatili di nuclei di comete un tempo attivi, che hanno perso tutti i loro gas nei ripetuti passaggi in prossimità del Sole. Se le cose stanno così, molti crateri terrestri e lunari sono di origine «cometaria», nel senso che le comete responsabili della loro formazione si erano trasformate precedentemente in oggetti Apollo, e pertanto sono già state considerate nei calcoli dei tassi di craterizzazione ad opera degli oggetti Apollo.

In qual modo gli oggetti Apollo capaci di formare grandi crateri sono correlati, se pure esiste una correlazione, con la miriade di corpi più piccoli, dalle grandi meteore fino alle particelle microscopiche, che bombardano continuamente l'atmosfera terrestre? Sembra probabile che questi corpi di dimensioni più ridotte provengano da vari tipi di sorgente. In breve si può dire, comunque, che una frazione notevole delle meteore osservate (distinte dalle meteoriti) non sono correlate a oggetti dall'aspetto di asteroide, ma derivano da comete osservabili. Le meteore cometarie sono quelle che si vedono

Le orbite della Terra e degli oggetti Apollo spesso si intersecano come due anelli di una catena. Dato però che le orbite degli oggetti Apollo subiscono una precessione piuttosto rapida, la loro orientazione rispetto all'orbita terrestre può cambiare significativamente in poche centinaia di anni. Il piano orbitale di 2100 Ra-Shalom è inclinato di 16 gradi (angolo i) rispetto al piano dell'orbita terrestre. L'illustrazione mostra le posizioni relative dei due corpi la notte in cui fu scoperto Ra-Shalom, il 10 settembre 1978. Le dimensioni e la forma dell'orbita ellittica di un asteroide sono precisate conoscendo la lunghezza del semiasse maggiore, a, e l'eccentricità dell'ellisse, e, definita come c/a, dove c è la distanza tra il Sole e il centro dell'elisse. L'orientazione del piano orbitale dell'asteroide rispetto a quello della Terra è dato da Ω, la longitudine del nodo ascendente, definita come la distanza angolare misurata verso est nel piano dell'orbita terrestre, cioè nel piano dell'eclittica, a partire

negli sciami periodici, come le Berseidi, che seguono l'orbita della cometa Swift-Tuttle, e le Draconidi, che sono associate alla cometa a breve periodo Giacobini-Zinner. Le meteore sono associate anche a comete attive di cui è stata osservata la disintegrazione e la scomparsa, come la cometa Biela, che diede origine alle Andromedidi. Pertanto, sembra ragionevole che almeno alcuni degli sciami meteorici in apparenza non correlati con comete conosciute siano in realtà resti di comete disintegratesi entro qualche migliaio di anni fa, mentre erano ancora attive.

D'altra parte, una frazione considerevole di meteore e bolidi non presenta evidenza di correlazione con sciami meteorici. Tra questi ci sono le meteoriti che sopravvivono al passaggio attraverso l'atmosfera terrestre. È molto improbabile che tutte le meteoriti che si trovano oggi nelle collezioni dei musei siano arrivate insieme a sciami, anche se alcune potevano essere resti di sciami cometari dispersi. Si può determinare il tempo di permanenza in orbita degli oggetti meteoritici misurandone il contenuto di prodotti di reazioni nucleari attivate dai raggi cosmici galattici. Si trova che quasi tutte le meteoriti sono state prossime per milioni di anni alle loro dimensioni attuali pari a meno di un metro. Anche se le meteoriti nacquero in sciami come resti di comete, nelle ultime migliaia di anni sarebbero state disperse in orbite casuali a seguito di interazioni gravitazionali coi pianeti.

Certamente, però, non tutte le meteore sono resti di comete. Per esempio, si può dimostrare che una percentuale piccola ma significativa di meteore, costituite presumibilmente da materiale abbastanza resistente da sopravvivere al passaggio attraverso l'atmosfera, deve provenire da frammenti prodotti in collisioni tra asteroidi ordinari avvenute nella fascia principale degli asteroidi tra Marte e Giove. A causa delle perturbazioni gravitazionali prodotte soprattutto da Giove, alcuni frammenti saranno stati spinti su orbite molto eccentriche che attraversano l'orbita terrestre. Qualsiasi frammento che capiti nell'atmosfera terrestre diventerà incandescente e apparirà come una meteora; quelli che sopravvivono e raggiungono il suolo saranno meteoriti. Naturalmente non c'è modo di identificare direttamente questi oggetti e di ritrovarne le origini nella fascia degli asteroidi per mezzo delle loro orbite.

Si può dimostrare qualitativamente come alcune meteore e meteoriti siano frammenti di oggetti Apollo. Almeno 19 oggetti Apollo hanno orbite che entrano nella fascia degli asteroidi o si estendono al di là di questa. Pertanto entreranno inevitabilmente in collisione con la grande quantità di detriti concentrata in quella parte del sistema solare. Oltre a varie centinaia di corpi di grandi dimensioni con diametri compresi tra 25 e 1000 chilometri si pensa che la regione contenga circa 10^{10} oggetti di dimensioni comprese tra pochi metri e un chilometro. Le collisioni tra questi piccoli asteroidi e gli oggetti Apollo devono essere frequenti.

In un incontro tipico l'oggetto più piccolo scaverà un cratere sull'oggetto Apol-

dall'equinozio di primavera e fino al punto in cui l'asteroide interseca il piano dell'eclittica da sud a nord. L'orientazione dell'ellisse dell'asteroide nel suo piano orbitale è indicata da ω, cioè dall'angolo tra il nodo ascendente e il punto del perielio, misurato nella direzione di movimento dell'asteroide. Nel caso di Ra-Shalom a vale 0,83 unità astronomiche (cioè 0,83 volte la distanza media tra la Terra e il Sole), e vale 0,436, Ω è 170,3 gradi e ω è 355,9 gradi. Le perturbazioni prodotte soprattutto dall'attrazione gravitazionale del pianeta Giove fanno compiere un giro completo (360 gradi) all'angolo ω in 10 000 anni circa. Lo schema a quadrifoglio a destra mostra che l'orbita dell'asteroide taglierà l'orbita terrestre in corrispondenza di quattro valori di ω separati da qualche migliaio di anni. Circa alla stessa velocità, la linea dei nodi (non rappresentata) precederà anch'essa di 360 gradi. Ogni 250 milioni di anni circa un oggetto Apollo tipico colliderà con la Terra.

lo di grandi dimensioni. La collisione impartirà una velocità di circa 50 metri al secondo al materiale espulso, che sfuggirà nello spazio, dato che la velocità di fuga da un corpo di un chilometro di diametro è molto bassa (minore di un metro al secondo). La velocità addizionale di 50 metri al secondo impartita ai detriti non sarà comunque sufficiente per cambiare apprezzabilmente la loro orbita rispetto a quella dell'oggetto Apollo stesso, che si muove intorno al Sole con una velocità di 20 chilometri al secondo circa. In qualche milione di anni i frammenti di collisione avranno buone possibilità di entrare nell'atmosfera terrestre.

Se un oggetto Apollo entra in collisione con un asteroide di 100 metri di diametro, c'è una probabilità elevata che l'oggetto Apollo stesso sia frantumato. La probabilità che lo sciame dei detriti colpisca la Terra è uguale a quella che aveva l'oggetto Apollo originario. Il tempo probabilmente necessario per un impatto sulla Terra sarà però molto più lungo di quello necessario per disperdere lo sciame di frammenti su orbite separate e praticamente casuali. Pertanto è molto più probabile che meteore e meteoriti prodotte nella craterizzazione o nella frantumazione di un oggetto Apollo arrivino singolarmente e sporadicamente piuttosto che come membri di uno sciame coerente. Naturalmente questo concorda con le osservazioni.

Ho cercato di quantificare la descrizione qualitativa che ho appena presentato per determinare quale percentuale del materiale meteorico e meteoritico, che si pensa bombardi la Terra in un anno, potrebbe essere spiegata con eventi collisionali che interessano gli oggetti Apollo nella fascia principale degli asteroidi. Secondo la mia valutazione circa 400 tonnellate di frammenti di oggetti Apollo dovrebbero entrare ogni anno nell'atmosfera terrestre sotto forma di corpi di peso compreso tra 100 grammi e 1000 chilogrammi. Se i frammenti hanno le proprietà meccaniche delle normali rocce meteoritiche, circa 100 delle 400 tonnellate dovrebbero sopravvivere al passaggio attraverso l'atmosfera, così da essere recuperabili come meteoriti. Anche se questo calcolo è incerto almeno per un fattore 10, il valore 100 tonnellate è molto prossimo alle stime correnti del tasso di impatto di meteoriti determinato in base al recupero effettivo di meteoriti, alle reti per l'osservazione fotografica delle meteore luminose e al tasso di impatti registrato dai sismografi lasciati sulla Luna.

A meno che il materiale degli oggetti Apollo sia particolarmente fragile e di facile distruzione nell'atmosfera terrestre, sembra praticamente certo che una parte notevole delle 2000 meteoriti che costituiscono le varie raccolte siano frammenti di oggetti Apollo, e può darsi che lo siano quasi tutte. La concorrenza da parte di altre sorgenti possibili non è forte. Ho già indicato che le meteoriti recuperate non appartengono a sciami associati a comete attive. L'unica sorgente alternativa richiede meccanismi capaci di perturbare resti di collisioni avvenute tra corpi della fascia principale degli asteroidi, spingendoli su orbite che attraversano l'orbita terrestre per mezzo di interazioni gravitazionali coi pianeti vicini, in particolare con Giove. È difficile calcolare la quantità di meteoriti che possono essere prodotte in questo modo, ma le stime più ragionevoli suggeriscono che questi meccanismi potrebbero fornire solo un decimo circa del materiale che bombarda la Terra, che può essere fornito dalla frammentazione di oggetti Apollo.

Se accettiamo che molte meteoriti, forse la maggior parte di esse, provengano da oggetti Apollo, che cosa ci possono dire le meteoriti sulla natura degli oggetti Apollo? L'85 per cento circa di tutte le meteoriti ritrovate appartiene alla classe di oggetti noti come condriti. Le condriti sono dette talvolta meteoriti non differenziate, dato che il loro contenuto di elementi meno volatili concorda praticamente coi rapporti di abbondanza degli elementi che si osservano nel Sole e nel complesso del sistema solare. In effetti, lo studio delle condriti ha fornito la maggior parte dei dati sull'abbondanza generale degli elementi e dei loro isotopi, l'età del sistema solare e le condizioni nel sistema solare primitivo.

Una prova sperimentale a favore della correlazione tra gli oggetti Apollo e le condriti fu ottenuta da Clark R. Chapman del Planetary Science Institute di Tucson, da Thomas B. McCord e Michael J. Gaffey dell'Università delle Hawaii e da Carle Pieters del Johnson Space Center di Houston, che misurarono la riflettività spettrale di 1685 Toro, il primo, tra i corpi che attraversano l'orbita terrestre, che abbia subito questo esame. La riflettività di questo oggetto Apollo in tutta la regione spettrale da una lunghezza d'onda di 0,3 micrometri nell'ultravioletto fino a 1,1 micrometri nell'infrarosso approssima strettamente la riflettività delle condriti tipiche. Dopo un picco a circa 0,7 micrometri entrambi gli spettri presentano una depressione in prossimità di 0,95 micrometri, prima di salire ancora. Nel caso delle condriti la depressione è dovuta a una grossa banda di assorbimento attribuibile ai pirosseni, minerali composti principalmente di silicati di magnesio e di ferro. Si può arguire che i pirosseni siano abbondanti anche in 1685 Toro.

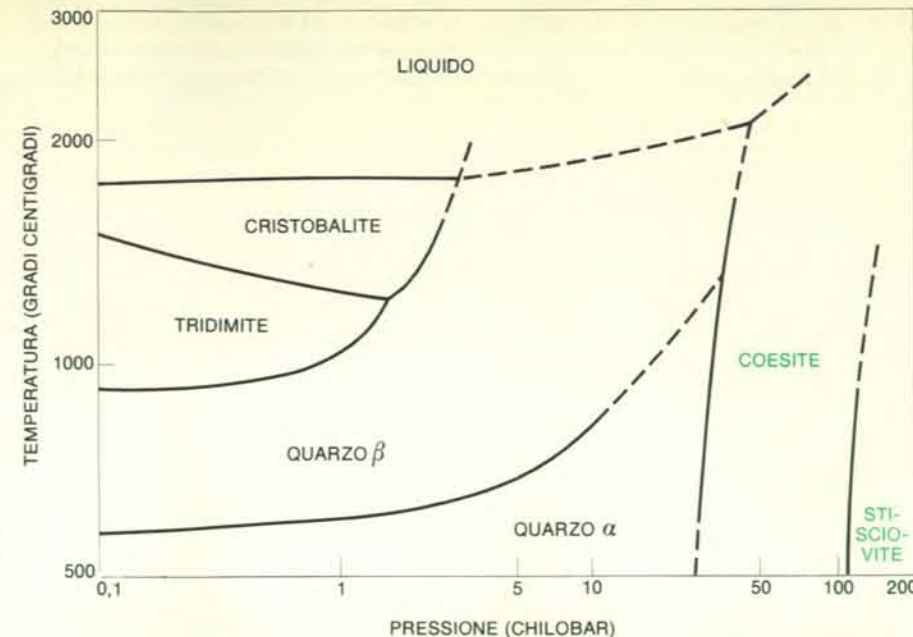
Le misure di albedo effettuate su altri oggetti Apollo sostengono l'ipotesi che siano di tipo condritico. Tre oggetti Apollo con bassi valori di albedo rivestono particolari interesse; due di questi sono gli ultimi scoperti: Ra-Shalom (il cui nome esprime una speranza di pace universale) e 1978 SB, appena scoperto all'Osservatorio astrofisico della Crimea. Le albedo di questi due oggetti sembrano abbastanza basse da assomigliare a quelle delle condriti carbonacee, il tipo di meteorite chimicamente più indifferenziato.

A questo punto possiamo chiederci: da dove vengono gli oggetti Apollo? Si può dimostrare facilmente che è errato credere al primo pensiero, in base al quale essi potrebbero esistere fin dalla formazione del sistema solare 4,5 miliardi di anni fa. Ho già fatto notare che la probabilità che un oggetto Apollo cada sulla Terra è pari a circa 5×10^{-9} per anno, e la probabilità di cadere su Venere è quasi la stessa. Pertanto, dopo varie centinaia di milioni di anni ci sarebbero pochi sopravvissuti nella popolazione degli oggetti che attraversano l'orbita terrestre. Nonostante questo, è possibile che la famiglia stimata attualmente di 750 oggetti Apollo rappresenti il resto di una popolazione inizialmente enorme? La risposta è no. Se fosse esistita una popolazione numerosissima, la registrazione dei crateri lunari ce lo avrebbe rivelato. Il flusso di proiettili capaci di scavare crateri è stato pressoché costante negli ultimi tre miliardi di anni. Se mai, c'è qualche risultato sperimentale che indica che la frequenza di formazione di crateri è aumentata leggermente negli ultimi 600 milioni di anni.

La riflettività spettrale (potere riflettente differenziale) dell'oggetto Apollo indicato come 1685 Toro *(barre verticali)* concorda con la riflettività di una normale condrite, il tipo più comune di pietra meteoritica *(curva continua)*. L'ampio avvallamento nella curva spettrale della meteorite nella regione attorno a 0,95 micrometri corrisponde a una forte banda di assorbimento dei pirosseni, che probabilmente sono i principali costituenti di 1685 Toro, l'unico oggetto Apollo studiato in dettaglio spettrale. Le misure furono fatte da C. R. Chapman del Planetary Science Institute, T. B. McCord e M. J. Gaffey dell'Università delle Hawaii e C. Pieters del Johnson Space Center.

È chiaro dunque che gli oggetti Apollo sono stati spinti nel sistema solare interno da qualche sorgente che opera con una frequenza quasi costante. In cifra tonda ogni milione di anni devono essere immessi circa 15 nuovi oggetti Apollo di diametro maggiore di un chilometro per bilanciare il tasso di perdita che si deduce ammettendo quattro collisioni con la Terra e tre con Venere ogni milione di anni. Nello stesso periodo un altro oggetto Apollo potrà entrare in collisione con qualche altro corpo (Mercurio, la Luna o Marte), e gli altri saranno spinti su orbite iperboliche e perciò espulsi definitivamente dal sistema solare. Sembrerebbe che ci siano due sole sorgenti possibili per i 15 oggetti Apollo richiesti. Essi possono essere piccoli asteroidi della fascia principale degli asteroidi deviati su orbite che attraversano l'orbita terrestre, oppure possono essere i residui privi di gas di comete un tempo attive. Oggi si comprende, almeno qualitativamente, come entrambe queste sorgenti possono assicurare un rifornimento di oggetti Apollo.

Piccoli asteroidi di dimensioni simili a quelle degli oggetti Apollo sono prodotti continuamente nella fascia principale degli asteroidi come frammenti di collisioni di corpi più grandi. I piccoli asteroidi saranno distrutti a loro volta da collisioni, così che il loro numero risulta stazionario. La popolazione stazionaria di asteroidi della fascia principale con diametro maggiore di un chilometro è pari a 400 000 unità circa. Si pensa che essi siano generati e distrutti dalle collisioni con un tasso di 400 circa ogni milione di anni. Un meccanismo capace di deviare il 4 per cento dei corpi appena prodotti su orbite che attraversano quella terrestre, basterebbe per conservare la popolazione degli oggetti Apollo.

Nonostante la piccolezza di questa frazione di asteroidi è molto improbabile che essi siano espulsi dalla fascia principale degli asteroidi e siano immessi in orbite che attraversano l'orbita terrestre a seguito diretto delle collisioni che li hanno generati. La variazione di velocità necessaria per trasferire materiale da un'orbita tipica di asteroide a un'orbita che attraversa quella terrestre è di circa sei chilometri al secondo. L'energia cinetica associata supera 10^{11} erg per grammo. L'energia necessaria per frantumare un corpo delle dimensioni di un oggetto Apollo non è superiore a 10^{8} erg per grammo. Pertanto, qualsiasi collisione capace di

La presenza di forme di quarzo ad alta densità (coesite e stisciovite) indica chiaramente che le rocce terrestri hanno subito pressioni elevate causate da un urto ad alta velocità. La silice presenta varie forme cristalline stabili in intervalli diversi di pressione e temperatura; le due forme di quarzo sono le più comuni. Le forme di alta temperatura, tridimite e cristobalite, sono conservate in forma metastabile dal rapido raffreddamento in rocce vulcaniche. La coesite e la stisciovite, che possono formarsi solamente a pressioni superiori a 30 chilobar (cioè 30 000 volte la pressione atmosferica), sono state trovate in prossimità di molti fra i crateri terrestri originati da impatto.

fornire l'energia cinetica necessaria per immettere un corpo delle dimensioni di un oggetto Apollo in un'orbita che attraversa quella terrestre, lascerebbe probabilmente frammenti di collisione molto piccoli. Inoltre, l'esperienza basata sull'osservazione di crateri terrestri da esplosione mostra che solo una parte molto piccola del materiale emesso raggiunge una velocità di sei chilometri al secondo; la maggior parte non raggiunge i 200 metri al secondo. Segue da ciò che anche se occasionalmente un frammento di grandi dimensioni potesse sopravvivere all'espulsione a sei chilometri al secondo, il numero di questi frammenti sarebbe comunque molto piccolo.

Pertanto si devono cercare processi meno violenti che possano determinare il trasferimento su orbite che intersecano quella terrestre. Se ne conoscono molti e tutti hanno in comune il fatto che il cambiamento di velocità richiesto è prodotto quasi completamente da perturbazioni gravitazionali. Il più importante di questi meccanismi comporta probabilmente vari tipi di interazioni risonanti capaci di aumentare l'eccentricità orbitale del piccolo corpo fino a farlo cadere sotto l'influenza gravitazionale di Marte.

A questo proposito può avere molta importanza il gruppo numeroso di asteroidi che si muove presso il limite interno della fascia degli asteroidi (da 2,1 a 2,6 UA dal Sole) su orbite che tagliano il piano dell'orbita di Giove con un angolo piccolo e sono completamente stabili su tempi dell'ordine dell'età del sistema solare. James G. Williams del Jet Propulsion Laboratory del California Institute of Technology ha dimostrato che le orbite di questi asteroidi hanno valori dei semiassi maggiori tali che la loro velocità di precessione è quasi risonante con quella del sistema accoppiato dei pianeti maggiori, dominato da Giove e Saturno. A causa dell'influsso dei loro giganteschi vicini gli asteroidi di questo gruppo particolare sono stati spinti su orbite che, nonostante la loro stabilità, oscillano considerevolmente in eccentricità.

I frammenti grandi come oggetti Apollo strappati a quegli asteroidi a seguito di collisioni accidentali non avranno però necessariamente la stessa stabilità. Se la velocità di un frammento fosse aumentata soltanto di 200 metri al secondo, e il frammento fosse posto in un'orbita con un semiasse maggiore poco più piccolo di quello del corpo da cui proviene, entrerebbe nella ragione critica in cui le risonanze col sistema accoppiato Giove-Saturno comincerebbero ad amplificare l'eccentricità orbitale, portandolo alla fine occasionalmente in prossimità di Marte, quando il pianeta è al suo afelio.

A causa di ripetuti passaggi in prossimità di Marte, le perturbazioni gravitazionali modificheranno casualmente il semiasse maggiore dell'oggetto Apollo potenziale, fino a una posizione in cui l'eccentricità orbitale dell'oggetto sarà amplificata da altre interazioni risonanti, che lo porteranno alla fine ad attraversare l'orbita terrestre, trasformandolo così in un oggetto Apollo vero e proprio. L'intero processo richiederebbe meno di 100 milioni di anni, e certamente alcuni oggetti Apollo sono diventati tali in questo

modo. La stima migliore indica, però, che questo processo potrebbe rendere conto soltanto di un oggetto Apollo circa ogni milione di anni. Bisogna stiracchiare molto le probabilità per spiegarne 15.

L'ipotesi alternativa, alla quale vanno i miei favori sulla base delle prove attuali, prevede che la maggior parte degli oggetti Apollo siano stati in precedenza delle comete. Si sa che le comete emettono volumi notevoli di materiale volatile, soprattutto vapore acqueo e anidride carbonica. Presumibilmente, all'inizio queste sostanze sono presenti nel nucleo delle comete sotto forma semplice di ghiaccio, oppure sono legate in composti più complessi. Queste sostanze volatili non sono però i soli costituenti del nucleo. Quando una cometa si avvicina al Sole si può osservare insieme coi gas anche molta polvere finemente granulata. Normalmente si liberano uguali quantità di polvere e di gas. Quando i frammenti più grandi di polvere non volatile sono catturati dalla Terra, producono gli sciami di meteore di origine cometaria.

Alcune meteore quasi certamente cometarie pesano varie tonnellate. Difficilmente oggetti così massicci possono essere trascinati con la scia usuale di gas volatilizzati; pertanto devono provenire da eventi più catastrofici, associati a scissioni del nucleo cometario, evento che si osserva occasionalmente. Senza un evento catastrofico di questo genere gli oggetti più grandi rimarrebbero semplicemente con la cometa, arricchendone il nucleo di materiale non volatile.

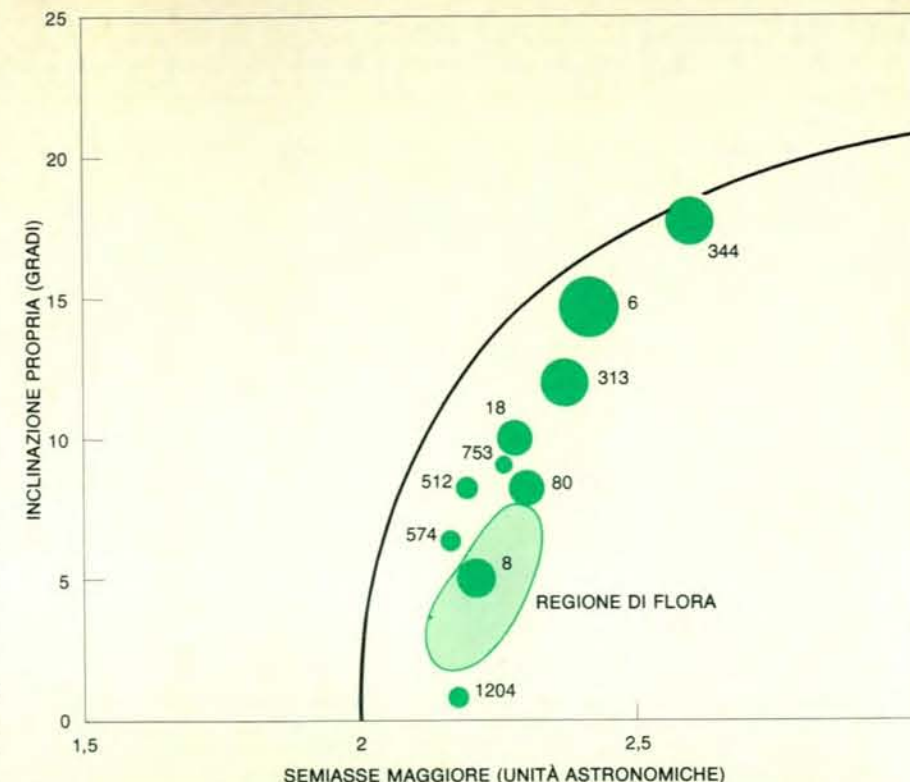
	OGGETTO	DISTANZA AL PERIELIO (UNITÀ ASTRONOMICHE)	DISTANZA ALL'AFELIO (UNITÀ ASTRONOMICHE)	INCLINAZIONE ORBITALE (GRADI)	DIAMETRO STIMATO (CHILOMETRI)	ANNO DELLA SCOPERTA
1	1566 ICARUS	0,19	1,97	23	1,0	1949
2	1978 SB	0,35	4,11	12	8,0	1978
3	1974 MA	0,42	3,09	38	4,0	1974
4	2101 ADONIS	0,44	3,30	1	0,8	1936
5	2100 RA-SHALOM	0,47	1,20	16	3,0	1978
6	1976 UA	0,47	1,22	6	0,2	1976
7	1864 DAEDALUS	0,56	2,36	22	2,0	1971
8	1865 CERBERUS	0,58	1,58	16	1,0	1971
9	HERMES	0,62	2,66	6	1,0	1937
10	1981 MIDAS	0,62	2,93	40	1,0	1973
11	1862 APOLLO	0,65	2,29	6	2,0	1932
12	2063 BACCHUS	0,70	1,45	9	1,0	1977
13	1685 TORO	0,77	1,96	9	4,0	1948
14	2062 ATEN	0,79	1,14	19	1,0	1976
15	1977 HA	0,79	2,40	23	0,6	1977
16	PL-6743	0,82	2,42	7	0,4	1960
17	1976 WA	0,82	3,03	23	2,0	1976
18	1620 GEOGRAPHOS	0,83	1,66	13	2,0	1951
19	1947 XC	0,83	3,67	1	2,0(?)	1947
20	1959 LM	0,83	1,84	3	6,0(?)	1959
21	1950 DA	0,84	2,53	12	2,0	1950
22	1866 SISYPHUS	0,87	2,92	41	5,0	1972
23	1973 NA	0,88	3,98	68	3,0	1973
24	1978 CA	0,88	1,37	26	2,0	1978
25	1863 ANTINOUS	0,89	3,63	18	2,0	1948
26	1975 YA	0,91	1,69	61	1,0	1975
27	PL-6344	0,94	4,21	5	0,4	1960
28	1978 DA	1,02	3,92	16	2,0	1978
29	1915 QUETZALCOATL	1,05	3,99	21	1,0	1953
30	1917 CUYO	1,06	3,23	24	2,0	1966
31	1580 BETULIA	1,12	3,27	52	6,0	1950

Ventotto sono gli oggetti Apollo noti oggi. Altri tre oggetti, gli ultimi tre della tabella, attualmente non attraversano propriamente l'orbita terrestre, ma la attraverseranno nel futuro a causa delle perturbazioni prodotte da Giove. Sedici dei 31 oggetti sono indicati con un numero: questo indica che i loro parametri orbitali sono noti con precisione sufficiente da poterli inserire nel catalogo degli asteroidi. I rimanenti sono indicati per lo più con l'anno della loro scoperta seguito da due lettere. La prima lettera indica la metà del mese dell'anno (iniziando con *A* per la prima metà di gennaio) in cui fu fatta la scoperta. La seconda lettera indica l'ordine di scoperta in quel periodo. Due oggetti sono indicati con la sigla PL, che sta per il catalogo Polmar-Leiden. Ra-Shalom dapprima fu chiamato 1978 RA, che significava che era il primo oggetto (A) scoperto nella prima metà di settembre (R) del 1978. Gli scopritori di un asteroide godono del privilegio di dargli un nome.

Le comete che appaiono per la prima volta hanno trascorso gran parte della loro vita viaggiando ben al di là del pianeta più lontano. Diventano luminose per la prima volta quando le perturbazioni prodotte da una stella vicina ne modificano la traiettoria, portandole su un'orbita con perielio prossimo al Sole. In molti casi le successive perturbazioni prodotte dai pianeti faranno uscire le comete dal sistema solare, prima che siano svuotate dei loro ghiacci volatili. Una piccola parte di queste comete a lungo periodo viene deviata, però, in orbite stabili più piccole, così che di solito passano presso il Sole ogni sette anni circa.

Nel corso di centinaia o migliaia di passaggi al perielio il nucleo della cometa viene svuotato sempre più delle sue sostanze volatili. Si è osservato che queste comete a breve periodo sono notevolmente meno attive delle comete che si muovono su orbite estese e molto eccentriche. In qualche caso le comete a breve periodo sono apparse completamente «asteroidali», nel senso che in alcuni passaggi al perielio non mostrano né chioma, né coda. Pertanto non occorre una grossa estrapolazione per immaginare che una cometa possa perdere tutte le sue sostanze volatili libere, così da assumere semplicemente l'aspetto di un asteroide.

La sola perdita delle sostanze volatili non basterebbe però per trasformare una cometa in un tipico oggetto Apollo. L'afelio di quasi tutte le comete a breve periodo si trova oltre l'orbita di Giove, mentre l'afelio di tutti gli oggetti Apollo conosciuti è interno a questa orbita. Ciononostante vi sono eccezioni. Varie comete a breve periodo hanno l'afelio appena all'interno dell'orbita di Giove. Una di esse, la cometa Encke, ha un afelio decisamente interno a 4,1 UA dal Sole.

La cometa Encke ha diminuito apprezzabilmente la sua attività in tempi storici, e sembra proprio sulla strada per diventare inattiva. Se dovesse diventare completamente inattiva, forse entro pochi secoli, sarebbe indistinguibile da un oggetto Apollo di piccola albedo. L'oggetto Apollo scoperto più recentemente, 1978 SB, è proprio un oggetto di questo tipo e la sua orbita è molto simile a quella della cometa Encke. È possibile che la cometa Encke e 1978 SB siano oggetti analoghi in uno stadio evolutivo diverso. L'esistenza di sciami meteorici in orbite a breve periodo analoghe, capaci di generare meteoriti con proprietà fisiche essenzialmente identiche a quelle delle meteoriti che si pensa siano generate dalla cometa Encke ci suggerisce che l'evoluzione orbitale di quella cometa non è un evento unico. Modelli fisici che possono spiegare una simile evoluzione orbitale sono stati presentati recentemente da Whipple e dal suo collega Zdenek Sekanina. Essi mostrano che si possono avere modifiche orbitali in una cometa in rotazione a seguito della reazione dinamica alla fuga di gas e polvere.

Una sorgente potenziale di oggetti Apollo è costituita da un gruppo di corpi ammassati in orbite di piccola inclinazione nella parte più interna della fascia principale degli asteroidi tra Marte e Giove. Questi asteroidi sono indicati qui col loro numero di catalogo; i cerchi pieni adiacenti ne mostrano le dimensioni relative. Si può valutare che il gruppo comprende circa 5000 oggetti *(area sfumata)* appartenenti alla regione di Flora (cosiddetta dal nome dell'oggetto più grande di questa regione: l'asteroide 8). Anche se tutti gli oggetti del gruppo si muovono su orbite stabili, sono vicini alla «superficie» di instabilità indicata dalla curva in nero. Nel 1969 James G. Williams, allora studente presso l'Università della California a Los Angeles, scoprì che i corpi portati dalla loro inclinazione e dal loro semiasse maggiore a giacere su questa superficie sarebbero perturbati da Giove e spinti su orbite molto eccentriche, che attraversano l'orbita terrestre. Quando gli asteroidi del gruppo qui illustrato collidono, può accadere che frammenti di dimensioni pari a quelle degli oggetti Apollo siano scagliati su orbite che sono sulla superficie di instabilità o prossime a essa.

Può sembrare che le prove fornite da una sola cometa, la cometa Encke, e da vari sciami meteorici cometari offrano solo un debole supporto a favore della origine cometaria degli oggetti Apollo. In realtà, anche nel caso che tutti gli oggetti Apollo fossero comete estinte, l'esistenza attuale di una sola cometa attiva come la cometa Encke può essere molto più di quanto ci si potrebbe attendere. Forse è solo un caso fortunato il fatto che la cometa Encke non è svanita qualche secolo fa; in tal caso non ci sarebbe una base osservazionale che ci spingerebbe a pensare che le comete si trasformano in oggetti Apollo, anche se fosse proprio così. Una sola cometa di tipo Encke ogni 65 000 anni fornirebbe oggetti Apollo al tasso di 15 ogni milione di anni, cioè al tasso necessario per mantenere la popolazione attuale in uno stato stazionario. Dato che la vita attiva di una cometa a breve periodo è solo di qualche migliaio di anni, per la maggior parte del tempo non ci sarebbero comete in fase attiva su orbite di tipo Encke.

Guardando al futuro, è possibile che si faccia un uso pratico degli oggetti Apollo. Alcuni di essi percorrono orbite così simili a quella della Terra da essere accessibili come la Luna. Inoltre, sono abbastanza piccoli che si può considerare la possibilità di usarli come stazioni spaziali già pronte e manovrabili. Secondo un'idea di Arthur C. Clarke valutata in dettaglio da Henry H. Kolm del Massachusetts Institute of Technology gli asteroidi verrebbero equipaggiati con «motori a massa», macchine che consumerebbero la materia dell'asteroide per spingerlo su nuove orbite. Altri hanno suggerito di estrarre dagli oggetti Apollo materie prime per la costruzione di veicoli spaziali.

Prima che si intraprendano programmi così fantasiosi potrebbe valer la pena di riflettere sul fatto che tutta la materia contenuta in tutti gli oggetti Apollo è solo di 10^{19} grammi circa, equivalenti soltanto ai primi due o tre centimetri di spessore della superficie continentale terrestre. È probabile che gli oggetti Apollo rappresentino, insieme con qualche cometa, la sola sorgente accessibile di acqua e composti di carbonio non terrestre nello spazio vicino, una sorgente la cui ricostituzione richiede circa 25 milioni di anni. Forse sarebbe meglio se la regione occupata dagli oggetti Apollo e dai loro parenti stretti, gli oggetti Amor, fosse dichiarata «regione protetta» inviolabile.

La dispersione di inquinanti in atmosfera a scala locale

Questo fenomeno, che dipende dal tipo di sorgente, dalla topografia del sito e dalle condizioni meteorologiche, non può essere studiato solo con normali metodi di laboratorio, ma richiede una filosofia ad hoc

di Domenico Anfossi, Flavio Bossa e Renzo Richiardone

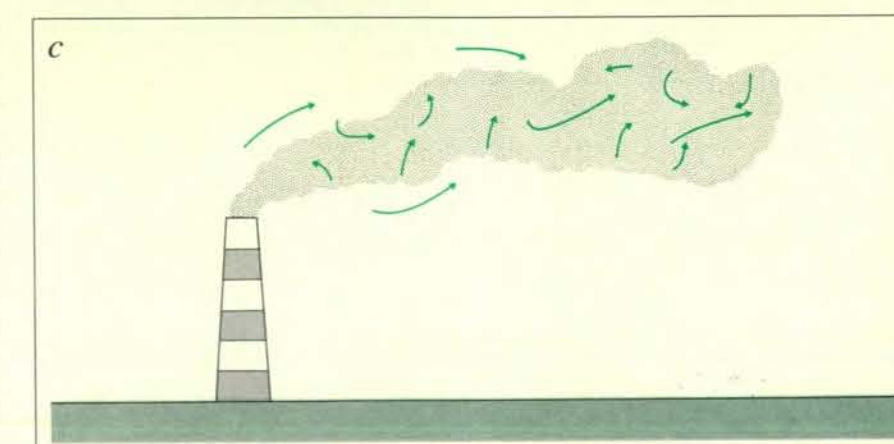
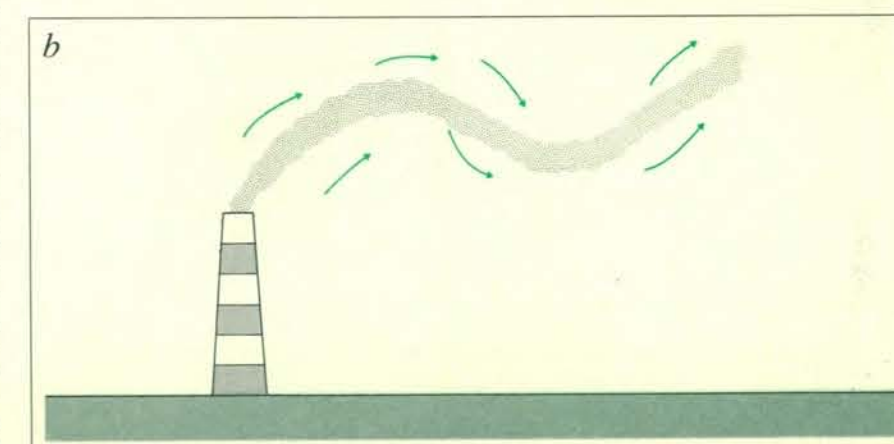
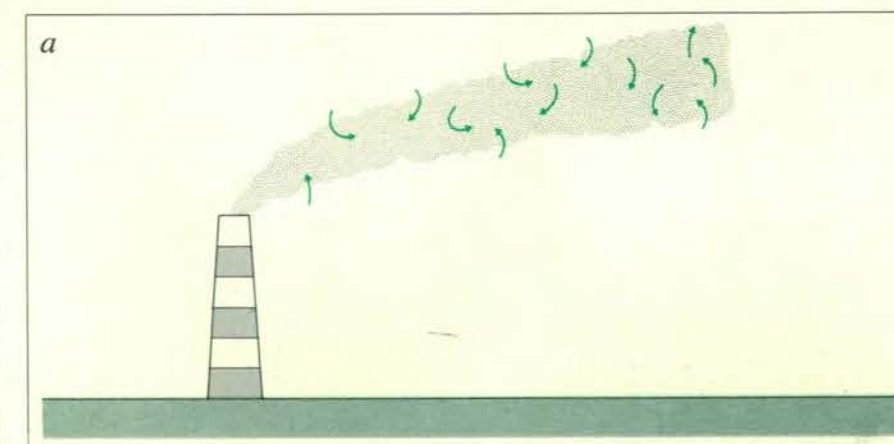
Il problema dell'inquinamento atmosferico, ovvero del deterioramento dell'«ambiente», nasce dall'imponente sviluppo raggiunto, a scala mondiale, dall'industrializzazione unito alla tendenza delle popolazioni a vivere concentrate in grosse città. Esso consiste nell'emissione in atmosfera di sostanze in concentrazioni elevate (o comunque superiori ai livelli presenti normalmente in natura) tali da produrre effetti dannosi non solo per la salute di uomini e animali, ma anche danni seri alla vegetazione e deterioramenti del suolo e dei materiali. Inoltre non vanno dimenticati gli effetti sul clima, la riduzione della visibilità e della radiazione solare e la creazione di situazioni di vera «calamità naturale» come il tristemente famoso disastro di Londra del 1952 dovuto allo smog, in cui si ebbero numerosi morti.

Sebbene il problema inquinamento sia emerso nella sua intierezza e drammaticità soltanto di recente, esso ha ovviamente origini lontane che si rifanno almeno a quando gli uomini scoprirono il fuoco. È interessante notare come il primo provvedimento legislativo «antinquinamento» risale al 1300 circa. Il re Edoardo I d'Inghilterra infatti, in tal epoca, emanò un editto che prevedeva la pena di morte per chi avesse bruciato carbone durante le sedute del Parlamento, e questo a causa del fumo e dell'odore prodotto. L'Inghilterra emanò dunque il primo editto, seguito da altri nel 1661, 1772 e da ben tre nel corso del 1800. Essendo stata però anche la prima nazione a vivere la Rivoluzione Industriale fu anche la nazione che per prima si trovò a dover affrontare questi problemi non solo con provvedimenti di legge ma anche con concrete azioni di ricerca (è del 1866 il primo lavoro scientifico sugli effetti dell'inquinamento sulla salute delle persone).

Comunque, come altre volte nella storia dell'umanità, i primi studi sulla diffusione in atmosfera, avviati con i dovuti mezzi e la necessaria profondità, furono determinati da esigenze belliche. Durante la prima guerra mondiale, furono usati intensivamente i gas (vedi per esempio l'iprite). Gli esiti però di tale nuova arma erano talvolta incerti se non addirittura avversi a causa dell'estrema variabilità del vento. Ciò indusse l'esercito inglese a indirizzare parte dei suoi ricercatori verso studi intensivi sulla diffusione dei gas in atmosfera. Risalgono così a quell'epoca le prime campagne di misura in atmosfera. Si ebbe comunque, come risultato, che la scuola inglese produsse negli anni fra le due guerre importantissimi risultati conservando anche nel primo dopoguerra la *leadership* in questo campo (si vedano ad esempio i lavori di G. I. Taylor, D. Brunt, O. G. Sutton e F. Pasquill). Fra le due guerre troviamo poi notevoli contributi dalla scuola tedesca (T. Von Karman e L. Prandtl) e da quella russa (A. N. Kolmogoroff, A. M. Obhukov, A. S. Monin e più recentemente D. L. Laitkman). Mentre la scuola inglese si è sviluppata lungo il filone empirico-statistico, la scuola tedesca e soprattutto quella russa, hanno cercato un maggior rigore logico-matematico. Negli Stati Uniti, il problema inquinamento atmosferico prende un notevole sviluppo solo dopo il 1943 in relazione alla costruzione e installazione dei primi reattori a energia atomica. Come uno dei risultati di tali studi, nel 1955 si ha la prima legislazione completa sull'argomento (che sarà poi aggiornata nel 1963 e nel 1970) seguita, nel 1956, da una analoga normativa in Inghilterra. La scuola americana si sviluppa lungo due correnti di ricerca diverse: la prima consiste nel trovare criteri o modalità empirici di controllo, misura e previsione dell'inquinamento, provati e il più possibile semplici, che siano basati sulle conoscenze già acquisite. La seconda linea tende invece ad approfondire lo studio dei fenomeni fisici che sono alla base della diffusione e del trasporto di inquinanti (si possono ricordare i nomi di D. B. Turner, F. A. Gifford, H. A. Panofsky e G. A. Briggs). Gli altri paesi europei (più il Giappone e l'Australia) si muovono anch'essi in maniera significativa nel dopoguerra. Oggi, studi in questo campo vengono effettuati non solo in tutti i paesi facenti parte da tempo della comunità scientifica internazionale, ma anche da paesi del terzo mondo, che insieme all'industrializzazione debbono anche affrontare i conseguenti problemi del deterioramento dell'ambiente. L'Italia è stato uno degli ultimi paesi a muoversi in questo tipo di ricerche, se si eccettuano lodevoli iniziative isolate al di fuori

TECNICA	TIPO DI ANALISI	SO_2	NO_x	CO	O_3	IT	ICM
Assorbimento UV	Misura dell'assorbimento della radiazione ultravioletta	X			X		
Chemiluminescenza	Misura fotometrica dell'emissione luminosa di reazioni chimiche (es.: O_3+NO_2)		X		X		
Calorimetria	Osservazione fotometrica delle variazioni cromatiche dei reagenti (per tecnica chimica umida)	X	X		X		
Conducimetria	Misura della conducibilità nelle soluzioni chimiche contenenti inquinanti	X					
Coulometria	Misura della corrente generata in particolari reazioni, proporzionale al tasso di assorbimento di inquinanti	X	X		X		
Elettrochimica	Misura della differenza di potenziale generata da un campione di aria inquinata, posto a reagire in una cella elettrochimica	X	X				
Ionizzazione in fiamma	Misura della corrente di ionizzazione in fiamma di un campione di gas per determinare il contenuto di C					X	
Fotometria in fiamma	Misura fotometrica della luminescenza di campioni di gas in fiamma da H_2	X	X				
SIND	Spettrofotometria della banda di assorbimento di un campione di gas di riferimento e correlazione con la banda di emissione di un gas incognito			X			
Gascromatografia	Metodo di analisi basato sulla migrazione differenziale di gas o solventi attraverso un mezzo assorbente poroso						X

In tabella sono elencati i principali tipi di analizzatori puntuali di inquinanti gassosi e le rispettive tecniche di analisi. Con SIND si indica la spettroscopia all'infrarosso non dispersiva; con ICM vengono indicati gli idrocarburi multicomponenti e con IT gli idrocarburi totali.

Il vento trasporta le molecole dei gas che vengono emessi nell'atmosfera o le particelle che vi si trovano in sospensione separandole in senso longitudinale in modo proporzionale alla sua velocità.

Se nell'atmosfera agissero solamente moti turbolenti a piccola scala il fumo si muoverebbe pressoché in linea retta *(a)*, con un graduale aumento della sua sezione trasversale; con moti a scala maggiore il pennacchio oscillerebbe invece ampiamente conservando una piccola sezione *(b)*. In realtà si osservano nell'atmosfera moti turbolenti con ampio spettro di scala, e ciò fa sì che il fumo *(c)* oscilli e si disperda nello stesso tempo. La distribuzione di tali scale non è uniforme.

L'andamento della temperatura con la quota influenza molto le capacità di rimescolamento verticale dell'atmosfera. Il profilo adiabatico (2) rappresenta in questo senso lo stato di equilibrio indifferente. Se la diminuzione di temperatura con la quota è superiore a 1°C ogni 100 metri il profilo è detto superadiabatico (1). Un gradiente di temperatura superadiabatico esalta i moti verticali dell'aria, mentre nel caso di profili isotermi (3) o di inversione (4) essi si smorzano.

dell'ambiente accademico. A riprova comunque di un cresciuto interesse nel nostro paese per questi studi, basterà ricordare il numero via via in aumento di istituti universitari o di laboratori del CNR che vi si indirizzano e l'apertura (1976), nell'ambito dei Progetti di ricerca finalizzati del CNR, di un canale di ricerca sull'ambiente (inteso nelle sue componenti generali: difesa del suolo, aria e acqua). In Italia la legge sull'inquinamento atmosferico è del 1966, mentre è del 1971 la pubblicazione sulla Gazzetta Ufficiale del regolamento per l'esecuzione della legge.

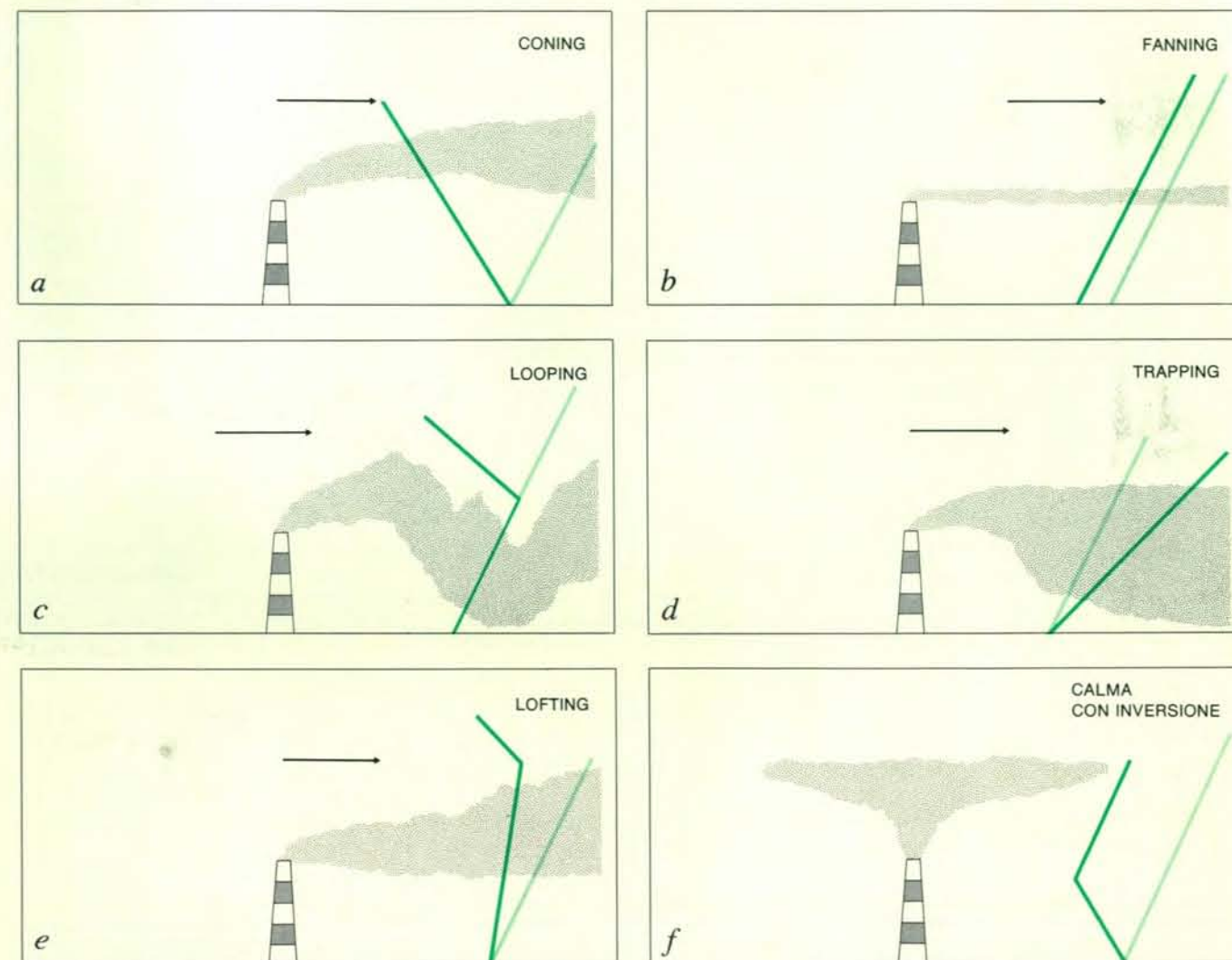
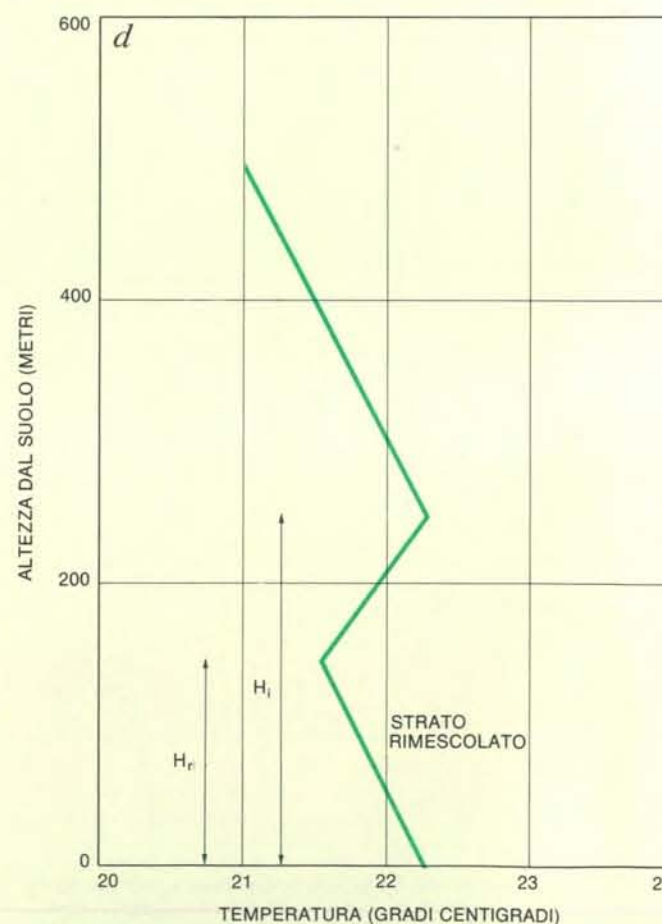
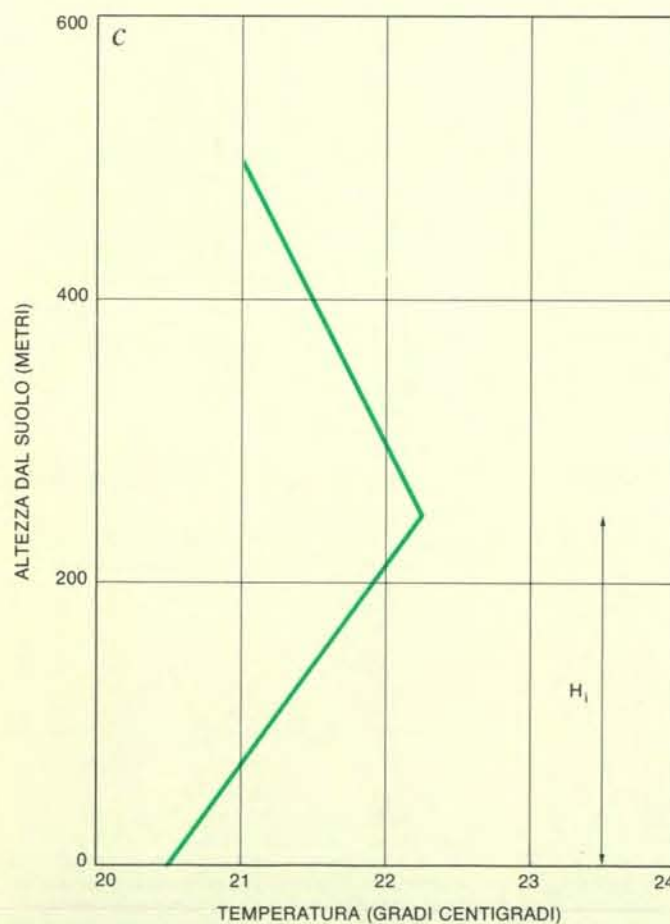
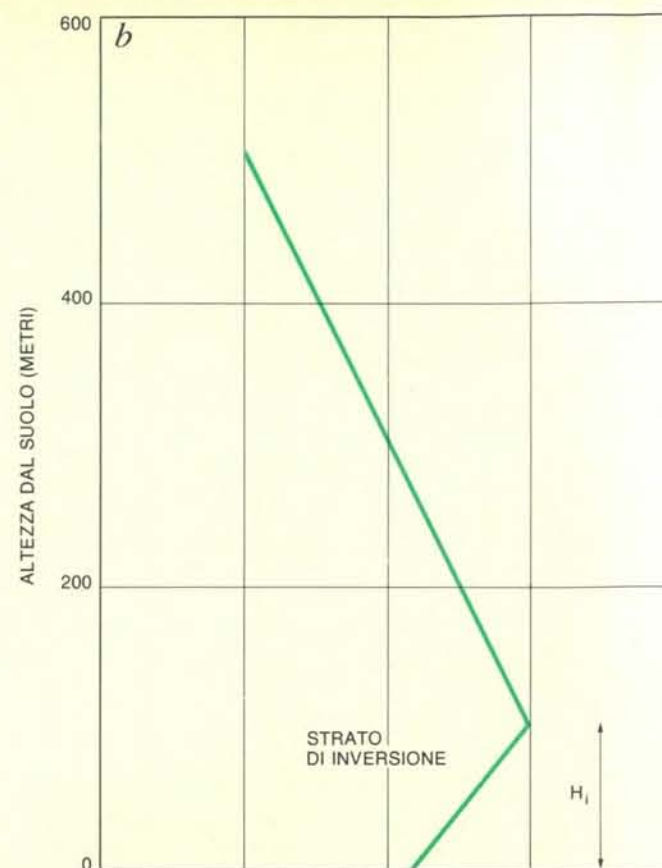
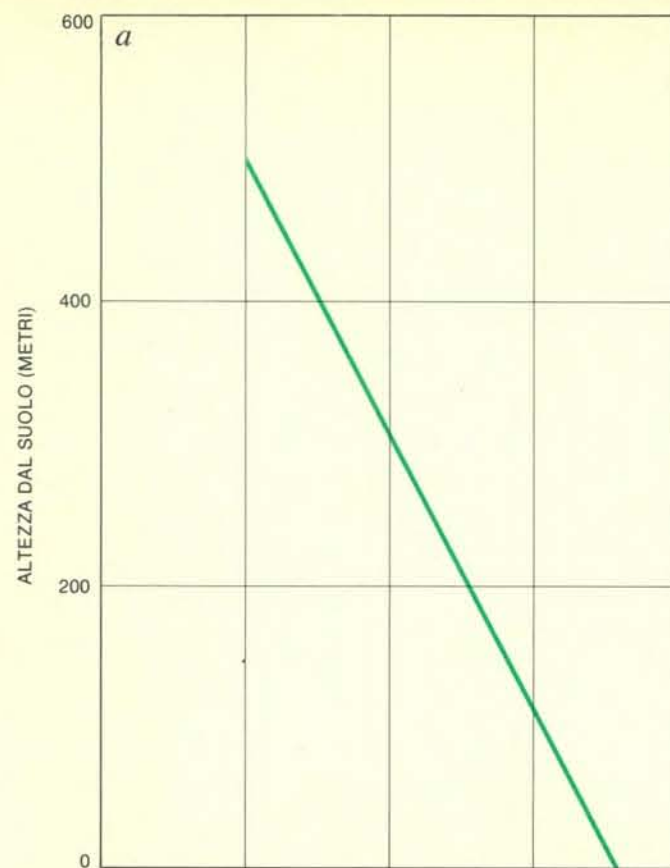
Schematicamente, si può dire che il problema inquinamento dell'aria come conseguenza dell'attività umana (che si sovrappone a quello normalmente presente in natura: eruzioni laviche, incendi di foreste, polveri trasportate dal vento, metano che si origina dalla decomposizione di sostanze organiche, altri idrocarburi emessi dalle piante, ecc.) si può scindere in tre sottoproblemi: emissione in atmosfera dai vari tipi di sorgente; trasporto e diffusione del materiale emesso (con relative trasformazioni chimiche subite) operati dall'atmosfera; interazione dell'inquinante con il suolo, con l'uomo, gli animali, le piante. La globalità del problema richiederebbe conoscenze dettagliate e aggiornate in campi di indagine fra loro anche molto lontani. Lo studio delle emissioni, infatti, richiede conoscenze di chimica delle combustioni e di ingegneria chimica e strutturale; lo studio del trasporto e delle reazioni chimiche che avvengono in atmosfera richiede la conoscenza di meteorologia generale, di micrometeorologia, di meccanica dei fluidi e di chimica; l'esame degli effetti dell'inquinante richiede infine conoscenze di medicina, fisiologia e biologia. Nel presente lavoro, verranno pertanto affrontati solo gli aspetti fisico-meteorologici del problema inquinamento con qualche accenno a misure di tipo chimico-meteorologico, intese sempre comunque come misure che concorrono a fornire un quadro, il più possibile generale, del fenomeno fisico oggetto di studio.

Una prima considerazione da farsi è legata alle scale dei fenomeni in studio, dato che i moti dell'aria interessano scale temporali che vanno dalla frazione del secondo alle decine di anni e scale spaziali che vanno dalla frazione del centimetro alla lunghezza del cerchio massimo terrestre. Ecco perciò che si parla di micrometeorologia (o meteorologia a scala locale) in cui si studiano i fenomeni entro una decina di chilometri e nell'intervallo di tempo dell'ordine dell'ora; mesometeorologia (cioè a scala regionale) in cui si arriva al centinaio di chilometri; meteorologia generale (cioè alla scala di una intera nazione o più). È comprensibile come ogni scala di fenomeni comporti teorie con ordini di dettaglio e di approssimazione diversi, così come abbastanza diverse saranno, in generale, le metodologie sperimentali; anche se, ovviamente, ci deve essere una possibilità di sovrapporre i risultati ottenuti alle varie scale.

Ci occuperemo qui della scala locale. In altri termini, il problema è il seguente: data una sorgente localizzata in un certo sito (pianeggiante, collinare, costiero, ecc.), posta a una certa quota sul piano campagna, emittente una certa quantità di inquinante nell'unità di tempo, studiare il destino di detto inquinante in relazione alla conformazione topografica e morfologica del terreno circostante e delle situazioni micrometeorologiche prevalenti e delle loro eventuali variazioni nel tempo. Va subito detto che una teoria completa ed esauriente, a tutt'oggi, non esiste, cosicché si è in grado di dare delle stime abbastanza attendibili solo in situazioni molto particolari come per esempio nei siti completamente pianeggianti e per di più interessati da venti di forte intensità. Si capisce come sia errato, per studi di previsione dell'impatto delle emissioni provenienti da un certo impianto sull'ambiente circostante, servirsi «acriticamente» dei risultati ottenuti in siti diversi da quello in esame. Si capisce anche come nasce l'esigenza di effettuare indagini, sito per sito, se si vuole raggiungere un livello di confidenza soddisfacente.

Quando ci si trova di fronte a un problema concreto si opera pertanto nel modo seguente: *a)* si procede a una descrizione qualitativa del fenomeno in oggetto sulla base dei dati sperimentali in possesso; *b)* si sviluppano dei modelli teorici (o più generalmente si utilizzano quelli già esistenti e provati in altri siti) in grado di descrivere quantitativamente il fenomeno; *c)* si progetta e si effettua una serie di campagne sperimentali in loco per valutare le prestazioni dei modelli mediante il confronto fra i dati previsti e quelli misurati, al fine di appurare quale dia i risultati migliori; *d)* si modifica eventualmente il valore dei vari coefficienti che tale modello conterrà in modo da migliorare l'accordo con le misure fatte (taratura del modello). Non va dimenticato come sia necessario e utile, anche se assai difficile, dedicare almeno parte delle energie disponibili all'approfondimento dello studio dei fenomeni fisici in sé, sia allo scopo di ridurre al minimo l'uso di coefficienti numerici dimensionali da valutarsi di volta in volta nei modelli, sia allo scopo di affinare la descrizione quantitativa dei fenomeni, riducendo i margini di incertezza.

Un'informazione necessaria all'effettuazione di stime di diffusione riguarda la configurazione della sorgente da cui gli effluenti (gas e/o particelle) vengono rilasciati nell'atmosfera. Il caso più semplice è costituito da una sorgente isolata e puntiforme (ciminiera); si possono poi avere più ciminiere a breve distanza l'una dall'altra (sorgenti multiple) o sorgenti che possono essere considerate lineari (una strada con traffico intenso o la scia di un aereo usato per irrorare campi coltivati); sorgenti areali (esempio tipico è una città con la sua moltitudine di sorgenti) e infine le sorgenti complesse che sono l'insieme di quelle viste precedentemente, quando si trovino l'una in presenza dell'altra. Le sorgenti isolate sono state fino a ora le più studiate a causa della loro maggior semplicità e importanza nel campo delle emissioni industriali e solamente in questi ultimi anni, con l'aumentata importanza dei problemi di inquinamento urbano, una maggiore attenzione è stata rivolta alle sorgenti areali.

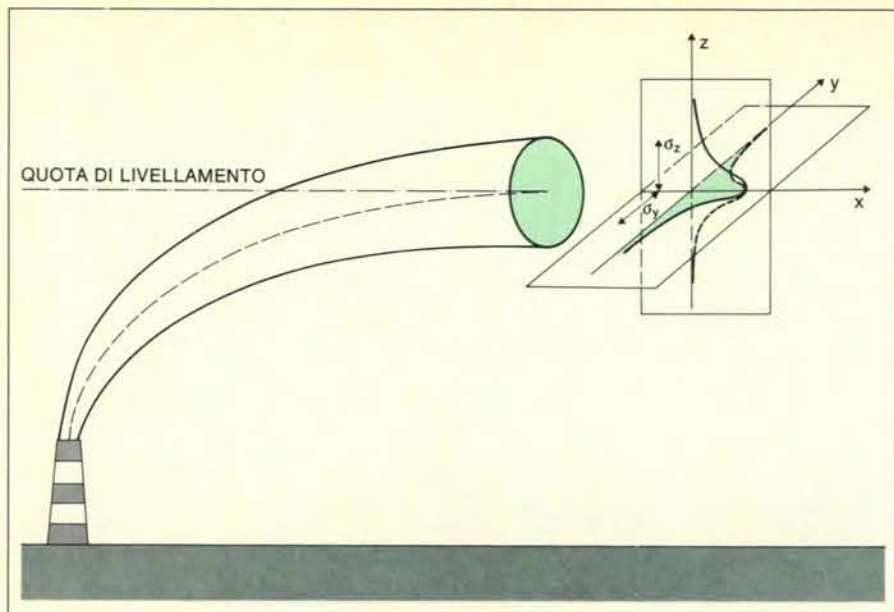
La dispersione dei gas emessi da una ciminiera è dovuta essenzialmente alla velocità del vento e alla sua turbolenza (che viene detta meccanica), a cui bisogna aggiungere gli effetti dovuti alla stratificazione termica verticale dell'aria. Queste grandezze non sono variabili indipendenti fra loro, ma vi è, più o meno esplicito, un legame di interdipendenza. Il vento trasporta le molecole del gas o le particelle in sospensione, appena lasciano la sorgente, separandole in senso longitudinale. La turbolenza meccanica (insieme di fluttuazioni di velocità del vento casuali, aperiodiche e scorrelate che si sovrappongono al moto medio dell'aria supposto quasi regolare) contribuisce invece alla diluizione laterale e verticale, rimescolando il pennacchio con l'aria circostante. La stratificazione termica infine può contribuire, come vedremo più avan-

L'esistenza di un legame tra i profili di temperatura e la forma di un pennacchio di fumo è all'origine della classificazione riportata in figura. Sebbene sia estremamente qualitativa, essa permette una schematizzazione delle capacità dispersive dell'atmosfera, tant'è che spesso la semplice osservazione di un pennacchio di fumo può fornire utili indicazioni sulle condizioni di stabilità dell'atmosfera. I profili di temperatura effettivi *(in colore intenso)* sono qui confrontati con i profili adiabatici *(in colore chiaro)*. Le situazioni più pericolose si hanno in presenza o di gradienti superadiabatici (quando il forte rimescolamento trascina il pennacchio al suolo provocando impennate dei valori di concentrazione) o di inversione. La presenza di uno strato di inversione in quota agisce da barriera per il fumo avente una quota di livellamento inferiore alla sommità dell'inversione stessa e pertanto esso si diffonde solo nello strato sottostante, causando situazioni critiche.

Nella sequenza dei quattro diagrammi è rappresentata schematicamente l'evoluzione diurna di un profilo di temperatura: dopo il tramonto *(a)* il raffreddamento del suolo si estende all'aria sovrastante causando la formazione di uno strato di inversione *(b)* che, nel corso della notte, raggiunge quote sempre più elevate *(c)*. Con il sorgere del Sole il riscaldamento del suolo si estende a uno strato atmosferico a forte rimescolamento che a poco a poco sostituisce lo strato di inversione *(d)* fino a riportare la situazione alle condizioni iniziali nelle ore più calde.

ti, ad amplificare, a smorzare e in alcuni casi anche a bloccare la dispersione della nuvola inquinante.

All'uscita dai camini, i fumi possiedono una certa velocità verticale e sono spesso a una temperatura molto più elevata di quella ambiente. Questi due fattori fanno sì che il fumo tenda inizialmente a innalzarsi in atmosfera, per poi piegarsi nella direzione del vento a mano a mano che il suo rimescolarsi con l'aria circostante ne diminuisce la temperatura. Nella prima fase di innalzamento, a questo processo contribuisce in maniera preponderante il regime altamente vorticoso proprio dei fumi all'uscita dalla bocca del camino (turbolenza autoindotta). Solamente quando la sua temperatura uguaglierà quella circostante, il pennacchio si troverà a essere completamente livellato; da questo punto in poi esso si comporterà in modo puramente passivo, seguendo abbastanza fedelmente i moti dell'aria (tranne le particelle di grosse dimensioni che precipitano, durante il trasporto, per effetto gravitazionale). La determinazione della quota di livellamento di un pennacchio di fumo è di particolare interesse, poiché le espressioni, che servono a calcolare le concentrazioni dell'inquinante al suolo, fanno dipendere queste ultime da tale altezza in ragione inversamente proporzionale al suo quadrato. Numerose sono state le formule proposte da vari autori per calcolare tale innalzamento, nessuna delle quali ha trovato una applicazione universale. E ciò è dovuto al fatto che tutte contengono qualche elemento di empiricità che va tarato di volta in volta. Si può comunque ritenere che i modelli sviluppati da Briggs (negli USA) sono quelli che mostrano una maggior adattabilità alle varie situazioni e sono quindi di più ampio impiego. Anche nei siti italiani che sono stati oggetto di indagine da parte del nostro gruppo, i modelli di Briggs si sono rivelati i più attendibili. Se la velocità di uscita e la temperatura dei fumi rispetto a quella ambiente sono trascurabili (è questo il caso più frequente nei camini delle abitazioni) essi possono essere trasportati al suolo dall'effetto aerodinamico di scia prodotto dall'edificio da cui fuoriescono; la pericolosità di questa situazione è evidente, poiché la concentrazione degli inquinanti è in questo caso elevata, a causa della vicinanza della sorgente.

Più sopra si è detto che un pennacchio è livellato quando la sua temperatura uguaglia quella dell'aria circostante (ovvero quando la turbolenza autoindotta del pennacchio cessa di essere prevalente e incomincia a predominare la turbolenza atmosferica). Se ne deduce che la temperatura dell'aria, e in particolar modo la sua variazione con la quota, hanno un ruolo fondamentale. Se un ipotetico volume d'aria si muove verticalmente nell'atmosfera, la sua temperatura varia a causa dell'espansione (o compressione) adiabatica che avviene nella fase di salita (o discesa). Tale variazione è di circa 1 grado centigrado ogni 100 metri (gradiente adiabatico secco). In uno strato atmosferico in cui vi sia un buon rimescolamento verticale delle masse d'aria (questo è il caso, per esempio, di venti di forte intensità) l'aria tenderà a essere caratterizzata da un gradiente verticale che è molto vicino a quello adiabatico secco. Tale tipo di profilo rappresenta, per i moti verticali nell'atmosfera, lo stato di equilibrio indifferente. Se infatti un certo volume d'aria possiede una temperatura diversa da quella circostante, esso tenderà a salire o a scendere a seconda che sia più caldo o più freddo dell'atmosfera in cui si muove. Ne segue che in un'atmosfera con un gradiente superadiabatico, ossia in cui la diminuizione di temperatura con la quota è superiore al valore di 1 grado centigrado ogni 100 metri, un volume d'aria, dotato di moto ascensionale, incontrerà nel suo cammino aria circostante più fredda, e quindi il suo moto verticale verrà rafforzato. Poiché anche un suo moto verso il basso verrebbe amplificato, a causa del minor riscaldamento del volume d'aria rispetto a quello dell'ambiente, un'atmosfera con un gradiente superadiabatico viene detta instabile. Quando si è in tali condizioni, il rimescolamento dell'aria è favorito, ragion per cui gli inquinanti eventualmente presenti saranno dispersi in volumi di aria assai grandi. Anche se questa sembra essere una situazione favorevole, bisogna ricordare che essendo queste condizioni accompagnate generalmente da venti di debole intensità (1 o 2 metri al secondo) le ricadute degli inquinanti si potranno avere a distanze brevi dalle sorgenti. Può succedere quindi che pur essendo molto diluiti nella direzione verticale, essendolo pochissimo in quella orizzontale, le concentrazioni risultanti al suolo siano particolarmente elevate. Al contrario, se il gradiente atmosferico è inferiore a quello adiabatico, l'atmosfera è detta stabile poiché i moti verticali vengono smorzati. In questo caso la dispersione verticale degli inquinanti in sospensione viene a essere notevolmente ridotta.

In un regime di circolazione tipico della brezza di mare i fumi emessi in uno strato stabile vengono fortemente dispersi solo quando incontrano lo strato instabile. Questo strato si forma per effetto del riscaldamento del suolo e la sua profondità aumenta con la distanza della linea di costa.

Negli strati atmosferici che si estendono dal suolo fino ad alcune centinaia di metri, si può osservare un vero e proprio ciclo giornaliero di profili di temperatura. Come tutti i corpi a temperatura maggiore dello zero assoluto, anche la superficie terrestre perde energia per irraggiamento (essendo la sua temperatura media intorno ai 300 kelvin, l'emissione radiante avviene nell'infrarosso). Durante il giorno, la perdita di energia sotto questa forma non è evidente dato che l'energia solare incidente è di intensità assai maggiore, ma nelle notti, specie in quelle in cui il cielo è sereno (le nuvole rappresentano una barriera per la radiazione infrarossa emessa dalla Terra e agiscono quindi da «specchio», riflettendole parzialmente verso il terreno) il suolo si raffredda, causando un raffreddamento degli strati d'aria a contatto con esso. Se il vento non possiede un'intensità sufficiente (6 o 7 metri al secondo) a rimescolare l'atmosfera, la diminuzione di temperatura nell'aria a contatto con il suolo dà origine a un profilo in cui la temperatura dell'aria aumenta con la quota, cioè a un profilo di inversione (inversione da irraggiamento). Nel corso della notte il raffreddamento interessa strati d'aria sempre più lontani dal suolo, cosicché la quota dell'inversione termica aumenta (secondo una relazione di proporzionalità con la radice quadrata del tempo), fino a raggiungere all'alba un massimo che, in base alle misurazioni effettuate dal nostro gruppo di ricerca, risulta nei mesi invernali, intorno ai 300 metri. Lo strato d'aria sotto la sommità dell'inversione è fortemente stabile, per cui la turbolenza atmosferica, e quindi la diffusione, sono gran-

Nelle applicazioni pratiche un pennacchio di fumo viene schematizzato supponendo che la concentrazione dell'inquinante vari in modo gaussiano intorno al suo asse. Le deviazioni standard nel piano orizzontale (σ_y) e verticale (σ_z) dipendono dalle condizioni di stabilità dell'atmosfera.

demente attenuate. La sommità dell'inversione stessa funge perciò da superficie riflettente per gli inquinanti che vi rimangono intrappolati (a meno che non posseggano un elevato eccesso di temperatura rispetto a quella ambiente, perché in questo caso, riescono a «passare» la barriera costituita dall'inversione).

Con il sorgere del Sole si verifica un fenomeno che può essere considerato l'opposto della situazione notturna. L'aria più a contatto del suolo si riscalda dando origine a un profilo superadiabatico, in cui la spiccata turbolenza garantisce un rapido e ampio rimescolamento. Il riscaldamento interessa via via strati d'aria sempre più alti e, quindi, aumenta la profondità dello strato rimescolato che viene a sostituire (dopo averlo distrutto) a poco a poco lo strato di inversione notturna (è la cosiddetta «erosione dell'inversione»). Gli inquinanti, che durante la notte si erano accumulati nello strato fortemente stabile, vengono rapidamente portati al suolo dai moti convettivi che si sviluppano entro lo strato rimescolato. Quest'ultimo fenomeno, che vien detto «fumigazione», può causare concentrazioni al suolo elevate e dura fino a che il limite superiore dello strato rimescolato non ha superato la quota raggiunta dall'inversione notturna (il che avviene, secondo le nostre misure, 3 o 4 ore dopo l'alba).

In realtà il ciclo diurno raramente si presenta in modo così schematico e semplificato, poiché la nuvolosità e il vento ne influenzano l'evoluzione. La presenza delle nuvole riduce infatti sia il riscaldamento diurno sia il raffreddamento notturno, mentre il vento, rimescolando l'atmosfera, tende a riportare i gradienti di temperatura al valore adiabatico. Durante i giorni nuvolosi e in presenza di vento di una certa intensità, si hanno infatti gradienti di temperatura pressoché adiabatici sia di giorno che di notte.

Anche le disomogeneità topografiche possono influenzare le caratteristiche dispersive dell'atmosfera. È stato riscontrato infatti da vari autori che le inversioni notturne avvengono con minor frequenza sulle città che non nei loro dintorni, poiché la città si comporta come una «isola di calore» sia a causa dell'energia liberata dalle attività umane, sia a causa dell'elevato potere assorbente dell'asfalto e delle case, che agiscono da volano termico liberando di notte l'energia (di origine solare) immagazzinata durante il giorno. In presenza di venti deboli si possono osservare dei moti ascendenti sulla città (aria più calda) accompagnati, per continuità, dalla discesa dell'aria in periferia. Tale tipo di circolazione può essere assimilato a una sorta di brezza (di campagna) analoga alle più note brezze di mare e di terra, o di monte e di valle, che hanno anch'esse origine dalla disomogeneità orizzontale di temperatura. Sostanze che in periferia sono confinate in uno strato ristretto a causa della forte stabilità, quando giungono sulla città, vengono disperse sia a causa di un gradiente molto meno stabile, sia a causa della turbolenza indotta dalla superficie accidentata del tessuto urbano. Un effetto simile si può verificare nelle zone costiere durante le ore diurne quando il maggior riscaldamento della superficie terrestre rispetto all'acqua dà origine al regime di brezza di mare (fenomeno che, come è noto, avviene in senso inverso di notte). Durante la brezza di mare, quando l'aria marina giunge sulla superficie terrestre che è più calda di quella del mare, si forma a contatto con il suolo uno strato instabile (caratterizzato da gradiente superadiabatico) che diventa sempre più profondo man mano che ci si inoltra verso l'entroterra. Anche in questo caso, i fumi emessi nello strato stabile verranno dispersi verso il suolo quando verranno a contatto, nell'entroterra, con lo strato instabile.

VELOCITÀ DEL VENTO AL SUOLO (m/s)	INSOLAZIONE DIURNA			CONDIZIONI NOTTURNE	
	FORTE	MODERATA	DEBOLE	COPERTURA ≥ 4/8	COPERTURA < 4/8
< 2	A	A-B	B		
2	A-B	B	C	E	F
4	B	B-C	C	D	E
6	C	C-D	D	D	D
> 6	C	D	D	D	D

A: CONDIZIONI ESTREMAMENTE INSTABILI
B: CONDIZIONI MODERATAMENTE INSTABILI
C: CONDIZIONI LEGGERMENTE INSTABILI
D: CONDIZIONI NEUTRE
E: CONDIZIONI LEGGERMENTE STABILI
F: CONDIZIONI MODERATAMENTE STABILI

Poiché dati meteorologici accurati sono raramenti disponibili, numerosi autori hanno suggerito la suddivisione delle caratteristiche turbolente dell'atmosfera in classi di stabilità, facilmente determinabili in base a semplici dati quali la velocità del vento al suolo, il grado di insolazione (di giorno) e di copertura (di notte). Per grado di copertura si intende la frazione di cielo coperta da nubi. La classificazione di Pasquill è una delle più adottate e si basa su 6 classi (dalla *A* alla *F*).

Anche se il gradiente termico verticale è il fattore che maggiormente influenza la turbolenza atmosferica nelle zone (quali ad esempio la pianura Padana) caratterizzate da frequenti calme di vento o comunque da venti di debole intensità, bisogna ricordare anche gli altri fattori che ne sono responsabili, vale a dire essenzialmente la rugosità del terreno e il profilo verticale del vento. Negli strati atmosferici superficiali le caratteristiche medie del vento (velocità e direzione) sono determinate essenzialmente dall'equilibrio di tre forze: il gradiente orizzontale di pressione, la forza di Coriolis e la forza di attrito viscoso dovuto alla presenza della superficie. L'azione di quest'ultima è massima ovviamente al suolo e si attenua via via che ci si innalza, per annullarsi approssimativamente intorno ai 1500, 2000 metri, per cui al di sopra di tale quota il vento può considerarsi con buona approssimazione dovuto solo all'equilibrio fra la forza di Coriolis e il gradiente orizzontale di pressione (tale vento viene chiamato geostrofico e la sua direzione è parallela alle isobare). Lo strato compreso fra il suolo e l'atmosfera libera (definita come zona atmosferica in cui si ha il vento geostrofico più sopra introdotto), è detto strato limite planetario. In esso, la presenza delle forze viscose, causando una diminuizione della velocità del vento sempre più intensa a mano a mano che ci si approssima al suolo, provoca anche una rotazione antioraria della direzione la cui entità dipende dalle condizioni di attrito («spirale di Ekman»).

Nell'atmosfera il ruolo della viscosità molecolare è trascurabile rispetto a quello della viscosità cosiddetta turbolenta. Mentre la prima è legata al trasporto (grazie all'agitazione molecolare) di quantità di moto tra zone di fluido che si muovono a diverse velocità, nella seconda tale trasporto è operato dai «vortici» turbolenti. Il termine vortice è in questo caso puramente simbolico e si riferisce in generale a quella varietà infinita di moti turbolenti, aventi scale diverse, che sono responsabili della diffusione in atmosfera di una qualunque grandezza che non vi sia distribuita in modo uniforme.

L'esistenza della spirale di Ekman è un buon esempio per comprendere come, in generale, le sostanze che vengono liberate in atmosfera non sempre si diffondono nella direzione del vento esistente a quota camino. In realtà il vento si comporta esattamente secondo la spirale di Ekman solo nei siti pianeggianti estesi (praterie). Gli effetti orografici producono distorsioni locali che si possono mettere in evidenza solamente con apposite misure. Le variazioni con la quota della direzione e della velocità del vento non influenzano solo le caratteristiche medie del trasporto di sostanze in sospensione (vale a dire la traiettoria media), ma anche quelle più legate alla turbolenza. Se infatti tali variazioni con la quota sono presenti, un volume d'aria che venga spostato da una quota all'altra giungerà alla nuova quota con una parte della sua quantità di moto iniziale e quindi costituirà una sorta di perturbazione rispetto al nuovo regime di moto che lo circonda. L'esistenza di una disomogeneità verticale nel profilo del vento tende pertanto a creare, o eventualmente a ingigantire, lo stato turbolento dell'atmosfera.

I coefficienti di dispersione laterale e verticale in funzione della distanza dalla sorgente a seconda della stabilità atmosferica, valutati da Pasquill in Inghilterra *(in colore intenso)*, sono posti a confronto con quelli rilevati in pianura Padana dal gruppo di ricerca congiunto Laboratorio di cosmogeofisica del CNR di Torino e dell'ENEL/CRTN di Milano, per evidenziare come la loro valutazione vari da sito a sito.

I primi approcci teorici al problema della diffusione atmosferica sono consistiti nel tentativo di applicarvi gli schemi e i concetti (quali il coefficente di diffusione, il libero cammino medio, l'equazione di diffusione) che si erano rivelati adatti a trattare la diffusione molecolare (sia di sostanze in sospensione che di calore, ecc.). L'introduzione del coefficiente di diffusione è legata all'ipotesi che, in presenza di un gradiente di concentrazione della quantità fisica che viene diffusa, il tasso locale di trasporto di tale quantità sia legato al gradiente da un fattore di proporzionalità che prende appunto il nome di coefficiente di diffusione. L'applicazione di tale modello all'atmosfera implica un'analogia tra i processi di agitazione molecolare e i vortici della turbolenza dell'aria. Tale ipotesi è fisicamente discutibile, ma ha portato a risultati incoraggianti in alcune situazioni. Mentre, infatti, nel caso molecolare gli effetti della diffusione si misurano su distanze che sono molto grandi rispetto alle scale tipiche del moto (libero cammino medio), ciò non avviene sempre in atmosfera (il continuo serpeggiamento di un pennacchio di fumo è un esempio evidente dell'azione di moti dispersivi che agiscono su scale maggiori della sezione del pennacchio). Nonostante questa limitazione di principio, la teoria del gradiente, a causa anche della sua semplicità formale, trova una grande applicazione presso vari autori. In particolare, è stata utilizzata con successo anche dal nostro gruppo per valutare il profilo termico verticale che si viene a creare di notte entro lo strato di inversione da irraggiamento.

Un altro metodo «molecolare» a essere applicato all'atmosfera è stato quello del moto browniano, in cui la trattazione del cammino delle molecole quale insieme di passi di uguale lunghezza distribuiti casualmente in ogni direzione ha condotto al risultato che la distanza media percorsa da una molecola è proporzionale alla radice quadrata del tempo trascorso dall'inizio del moto. L'applicazione di tale metodo all'atmosfera non può prescindere dal fatto che in essa agiscono vortici di scala diversa, e con diversa frequenza statistica. Appare necessario tener conto del coefficiente di correlazione tra le fluttuazioni del vento a istanti diversi, come suggerito da Taylor nel 1921, e ciò conduce a una dipendenza della dispersione dalla radice quadrata del tempo solo per tempi lunghi (se confrontati con le scale tipiche dei moti turbolenti) mentre per tempi brevi la dispersione è direttamente proporzionale al tempo. L'acquisizione di dati statistici sulle fluttuazioni di velocità a cui è sottoposto un piccolo volume di aria nel corso del suo moto (statistica detta lagrangiana) viene effettuata con l'uso di traccianti. A differenza di quanto avviene per le registrazioni di fluttuazioni della velocità del vento in un punto fisso mediante anemometri (statistica detta euleriana), le misure con i traccianti sono necessariamente intermittenti, e non si possono svolgere in modo completamente automatizzato. Nasce pertanto l'esigenza di trovare un legame tra questi due tipi di statistica. J. S. Hay e Pasquill hanno suggerito che, affinché le due rappresentazioni conducano allo stesso valore di dispersione, è necessario che i due tempi caratteristici di scala (euleriano e lagrangiano) siano fra loro proporzionali. Dovendo fare delle operazioni di media per giungere a stime di dispersione di materiale in sospensione, quanto più sopra esposto significa che per ottenere da misure anemometriche lo stesso risultato che si sarebbe ottenuto da misure di traccianti, per le stesse distanze sottovento, basterà operare con un tempo di media ridotto di un fattore che tenga conto del rapporto fra le due scale, rispetto a quello che si sarebbe usato nel caso lagrangiano.

In figura è riportato lo schema di funzionamento del lidar (acronimo di *Light Detection And Ranging*). L'emissione degli impulsi, della durata di alcuni nanosecondi, con alta potenza di picco, avviene per mezzo di un interrutore ottico *(Pockels-cell)* che, se chiuso, interdice l'emissione laser dando modo alla inversione delle popolazioni dei livelli energetici di raggiungere valori estremamente elevati; la sua apertura istantanea innesca l'azione laser in condizioni che consentono amplificazioni d'onda estremamente elevate. La cella di Pockels ha il compito di lasciar passare la radiazione polarizzata attraverso un cristallo anisotropo al quale sia stata applicata una tensione di alcune migliaia di volt (5 ÷ 6 kV). Quando l'energia trasmessa dal laser passa attraverso l'atmosfera, le molecole di gas e le particelle incontrate retrodiffondono parte della radiazione incidente in direzione del lidar dove viene raccolta da un telescopio in configurazione di Cassegrain e trasmessa a un fotomoltiplicatore e di qui, attraverso una catena di amplificazione, viene mandata su una unità di ricezione e di presentazione dati (segnale analogico). Il filtro a interferenza, posto prima del fotomoltiplicatore, serve a tagliare il più possibile la radiazione a larga banda del Sole.

Poiché nelle applicazioni pratiche appare desiderabile poter fornire stime di diffusione in base a dati meteorologici facilmente disponibili (e che pertanto non comportino misure dettagliate di fluttuazioni del vento), oppure effettuare previsioni di massima in località non ancora studiate da un punto di vista meteorologico, sono state suggerite da numerosi autori alcune schematizzazioni dei processi diffusivi, in cui la concentrazione dell'inquinante attorno alla sua quota di livellamento è supposta essere distribuita in modo gaussiano sia sul piano verticale sia su quello orizzontale. Tale assunzione, anche se non è pienamente giustificata, si è rivelata in discreto accordo con i dati sperimentali, specialmente quando essi riguardano valori mediati su intervalli di tempo dell'ordine della mezz'ora o più. Essa ha inoltre notevoli vantaggi di natura statistica, primo fra i quali è quello di far sì che la forma della distribuzione sia completamente determinata dai suoi due primi momenti: valore medio e deviazione standard. In particolare, si definiscono due diverse deviazioni standard (σ_z e σ_y) rispettivamente in riferimento al piano verticale e orizzontale. Non si tiene conto della deviazione standard lungo la direzione x (che è la direzione del vento medio), in quanto il termine di trasporto (proporzionale alla velocità del vento) prevale su quello diffusivo (dovuto alla turbolenza). Questa ipotesi cessa però di valere per velocità del vento molto basse. σ_z e σ_y saranno funzione, fra l'altro, delle condizioni meteorologiche esistenti e, a tal proposito, numerosi metodi empirici sono stati proposti per la loro valutazione. Alla base di tali metodi vi è la suddivisione delle caratteristiche dispersive dell'atmosfera in classi di stabilità, in corrispondenza delle quali vengono definite espressioni di σ_z e σ_y in funzione della distanza sottovento che sono state ricavate da dati sperimentali. La stabilità viene determinata in base a dati meteorologici essenziali e facilmente rilevabili, quali l'irraggiamento solare, lo stato di copertura del cielo e la velocità del vento al suolo. La prima di tali classificazioni, e ancora oggi la più usata, è quella dovuta a Pasquill.

Questi cenni alla fenomenologia e alle teorie fondamentali della dispersione di inquinanti in atmosfera evidenziano, a nostro parere, la complessità di questi tipi di studio, specialmente in relazione all'estrema variabilità delle grandezze in gioco. Se alle considerazioni suddette si aggiungono le complicazioni ulteriori prodotte dalla struttura orografica e morfologica dei siti «reali» in cui gli

VARIABILI METEOROLOGICHE	SENSORI DIRETTI	PRINCIPIO DI FUNZIONAMENTO
Velocità e direzione del vento al suolo o a quote fisse	Anemometri meccanici a coppette o a elica (ventola)	Trasduzione elettrica di un segnale meccanico
Fluttuazioni turbolente della velocità e direzione (misure di microturbolenza)	Anemometri triassiali meccanici	Trasduzione elettrica di un segnale meccanico
	Anemometri a filo caldo o a film caldo (a corrente o temperatura costante)	Variazioni della resistenza dell'elemento sensibile (filo o film caldo) in funzione della velocità del vento
	Anemometri sonici	Dipendenza del tempo di transito di un impulso acustico dalle caratteristiche del mezzo in cui si propaga (aria)
Profili verticali di velocità e direzione del vento e gradienti di velocità	Anemometri dei vari tipi posti su torri meteorologiche o lungo camini o su alture a quote diverse	Trasduzione elettrica di un segnale meccanico
	Palloncini Pilot	Inseguimento mediante teodoliti aerologici o telemetri laser o radar di un palloncino con spinta ascensionale costante
	Palloni sonda	Inseguimento via radar o con tecniche radio
Traiettorie dell'aria	Palloncini equilibrati	Inseguimento mediante teodoliti aerologici di un palloncino equilibrato (spinta ascensionale nulla)
	Tetroon	Inseguimento via radar o radio (i tetroon sono palloni di forma tetraedrica a volume costante)
Temperatura	Termometri a mercurio, termocoppie, termoresistenze	Trasduzione elettrica o meccanica
Profili verticali di temperatura	Termometri montati su torre, radiosonde appese a palloni frenati o liberi o aerotrasportate	Trasduzione via radio o via cavo di segnali elettrici
Pressione	Barometri, barografi, trasduttori meccanici	Trasduzione elettrica di segnali meccanici
Umidità relativa e punto di rugiada	Termometri a bulbo bagnato, psicrometri, igrometri, strips Li-Cl e altri composti chimici	Trasduzione elettrica o meccanica
Radiazione solare	Solarimetri, piranografi	Misura della radiazione diretta e diffusa
Radiazione netta	Net-radiometri	Misura della radiazione netta mediante termopila
Precipitazioni	Pluviometri, nivometri	Pesatura del precipitato e misura del livello

Gli strumenti meteorologici possono essere suddivisi in sensori a misura diretta e in sensori a telerilevamento *(si veda la tabella alla pagina successiva)*. I primi qui elencati sono quelli che misurano direttamente le grandezze fisiche del mezzo (atmosfera) in cui essi stessi sono posti.

VARIABILI METEOROLOGICHE O CHIMICHE	SENSORI A TELERILEVAMENTO	TIPO	PRINCIPIO DI FUNZIONAMENTO
Temperatura e stabilità atmosferica, spessore dello strato rimescolato	Lidar (vari tipi, sia bistatici che monostatici)	A, El	Analisi della retrodiffusione delle particelle
	Radiometro a microonde	A, El	Emissione nella banda 50-60 GHz dell'O_2
	Sodar	A, Ac	Retrodiffusione dovuta alle fluttuazioni termiche
	Rass	A, El, Ac	Analisi di un'onda acustica per mezzo di un radar Doppler
	Radar	A, El	Analisi di due impulsi mediante effetto Doppler
Emissività del terreno	Radiometro all'infrarosso	P, El	Aviotrasportato, misura l'emissione del suolo nella banda IR
Visibilità, nebbia	Trasmissometro	A, El	Attenuazione del fascio di luce in atmosfera dovuto ad assorbimento
Altezza nubi	Telefotometri, fotometri a Sole	P, El	Attenuazione modulata di frequenze tipiche di emissione del Sole e analisi per confronto fra radiazione solare e radiazione di emissione
	Nefelometro	A, El	Retrodiffusione atmosferica
Umidità	Igrometri all'infrarosso	A, El	Misura di assorbimento della radiazione infrarossa per confronto con la banda di assorbimento del vapor d'acqua (1,37 μm)
Velocità e direzione del vento	Anemometro laser (laser Doppler)	A, El	Misure di ritardo, su un fascio laser continuo, per effetto Doppler
	Telemetria laser		Teleinseguimento
Misura integrata di concentrazione al suolo di inquinanti tossici (SO_2, NO_2)	Spettrofotometro a correlazione di maschera	A, P, El	Confronto di concentrazioni di SO_2 e NO_2 mediante analisi spettrale
Dimensioni e struttura fisico-chimica dei pennacchi di fumo	Lidar con spettrofotometro a correlazione di maschera	A, El	Analisi della retrodiffusione delle particelle più analisi spettrale delle concentrazioni di inquinanti gassosi

I sensori a telerilevamento consentono misure a distanza perché basano il loro funzionamento sull'analisi dell'effetto del mezzo su un'onda elettromagnetica (El) o acustica (Ac) che vi si propaga. Nel caso in cui tali sensori utilizzino onde emesse dall'ambiente, essi sono chiamati passivi (P), mentre nel caso in cui le onde siano emesse artificialmente si dicono attivi (A).

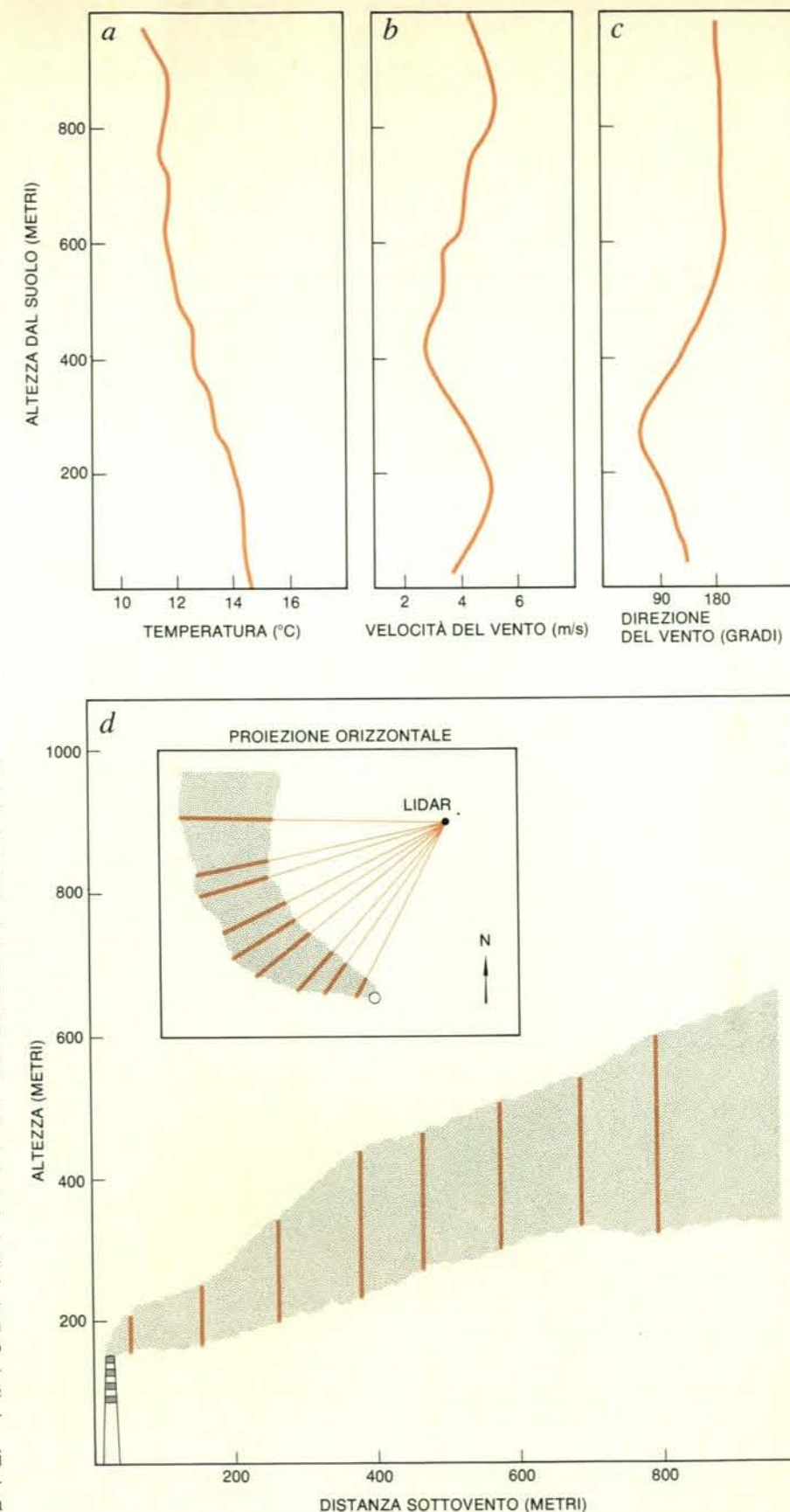
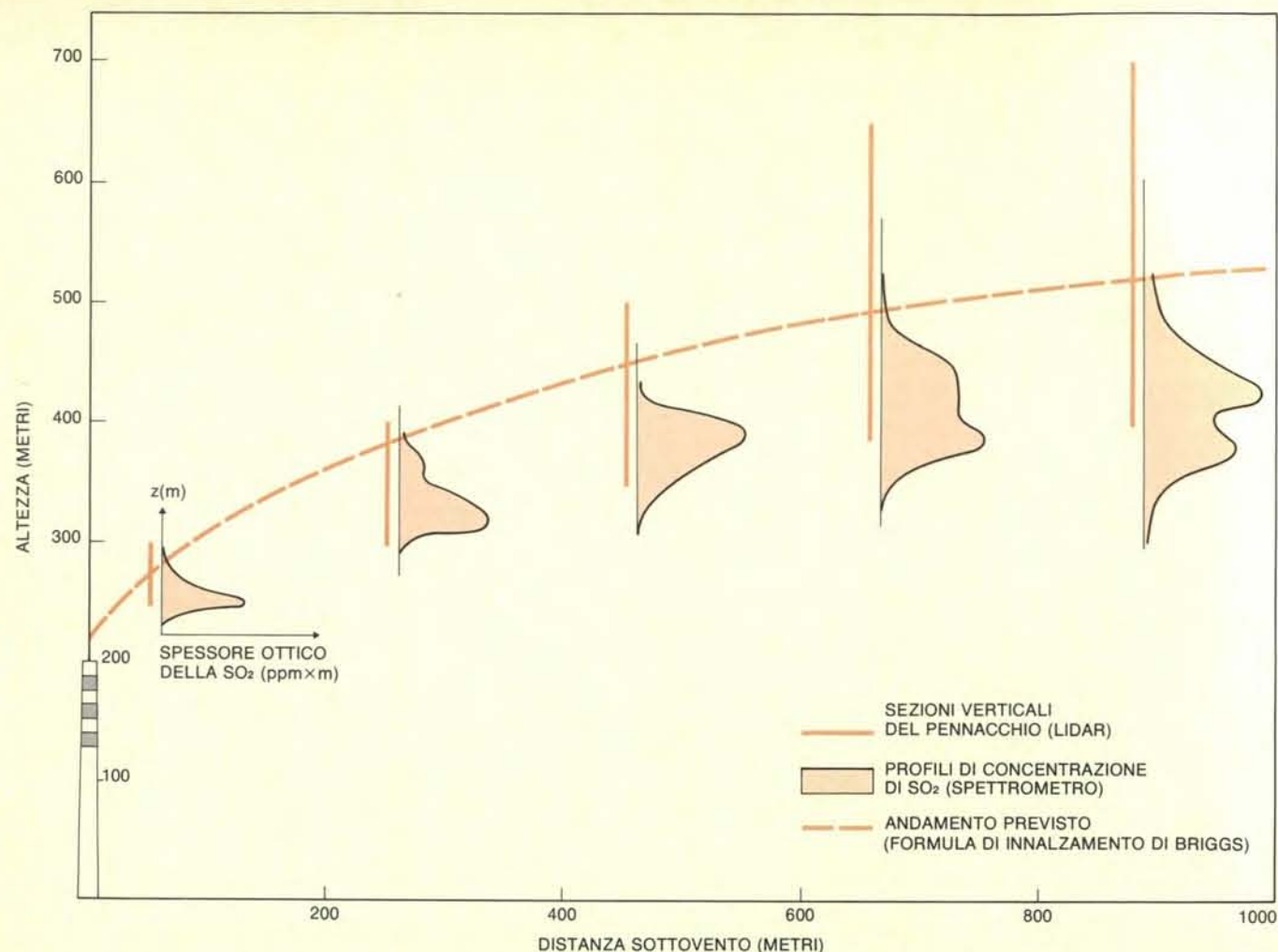
impianti industriali sono stati costruiti o in cui si progetta di installarli, si capisce come le metodologie sperimentali debbano, di necessità, essere molteplici e complesse.

Un'indagine sperimentale che voglia conseguire risultati attendibili deve pertanto essere condotta disponendo, in opportuni punti del sito in esame, strumenti in grado di misurare contemporaneamente i parametri di interesse e le loro variazioni sia nel tempo che nello spazio. Mentre le misure al livello del suolo possono essere pressoché automatizzate, richiedendo quindi solo controlli saltuari da parte degli operatori, le misure in quota (che sono le più importanti per quanto detto) richiedono quasi sempre la presenza continua di personale. Ciò limita di fatto la durata e il numero delle misurazioni effettuabili, rendendo quindi impossibile eseguire le misurazioni col dettaglio spazio-temporale che sarebbe necessario. In particolare, questo fatto costituisce una limitazione al contributo che la sperimentazione porta alla comprensione fisica dei processi diffusivi. Ciò nonostante, è solo acquisendo nuove conoscenze tramite l'osservazione e la misurazione in campo dei fenomeni che si può cercare di colmare il divario conoscitivo che si viene a creare fra i modelli proposti per valutare quantitativamente i livelli di concentrazione al suolo e i risultati delle misurazioni di detti valori, ottenuti, in genere, con reti fisse di monitoraggio costituite da prelevatori puntuali. Detti prelevatori, che funzionano in modo completamente automatizzato, analizzano campioni d'aria e forniscono i valori di concentrazione mediati su un certo intervallo di tempo (solitamente mezz'ora). Essi hanno ormai incontrato una certa diffusione commerciale presso gli enti pubblici (amministrazioni locali) e le grosse industrie. Vengono dislocati in modo da formare una rete di monitoraggio atta a segnalare tempestivamente situazioni particolarmente pericolose.

I prelevatori puntuali sono indubbiamente gli strumenti più adatti a segnalare il superamento di valori critici di concentrazione in un punto di particolare interesse, ma mostrano il fianco ad alcune critiche. Se da un lato, infatti, è utile conoscere quando in una certa zona si superano i limiti di tollerabilità di un certo inquinante, questa informazione diviene poi pressoché inutile se non si sa su quali impianti intervenire per ridurre le concentrazioni al suolo. In altri termini, come è ovvio del resto, una rete di monitoraggio non è in grado, da sola, di discriminare fra tutte le sorgenti emettenti in una certa area, quelle che contribuiscono maggiormente, in quella situazione particolare, ad aumentare i livelli di concentrazione misurati da un certo prelevatore. Dovendo necessariamente limitare il numero dei misuratori, sorge inoltre l'esigenza di ubicarli nei luoghi più significativi, vale a dire in quelle zone ove si possono riscontrare i livelli di inquinamento più elevati. Da tutto ciò deriva, come già ricordato, la necessità di una conoscenza la più possibile accurata della situazione meteorologica locale e della messa a punto di un modello preditttivo il più possibile aderente alla realtà.

Gli strumenti oggi in commercio che possono utilmente essere impiegati durante una campagna di misurazioni sono troppi perché un esame dettagliato delle loro peculiarità e prestazioni possa essere affrontato in questa sede. Per una loro panoramica si vedano la tabella a pagina 33 e la tabella nella pagina a fronte; qui ci occuperemo di quelli che, vuoi per l'importanza della grandezza fisica misurata, vuoi per i principi su cui si basano, sembrano essere di maggior interesse. Recentemente l'attività sperimentale in campo micrometeorologico ha potuto compiere un notevole salto sia qualitativo sia quantitativo grazie all'estendersi delle tecniche di misurazione a distanza (telerilevamento o *remote sensing*). A differenza delle tecniche più tradizionali in cui la strumentazione impiegata viene utilizzata per misurare le grandezze fisico-chimiche del mezzo in cui si trova lo strumento (si parlerà in questo caso di strumenti a misura diretta), uno strumento di remote sensing ne consente indirettamente la misura a distanza poiché analizza l'effetto da esso esercitato sulle caratteristiche di propagazione di una onda elettromagnetica o acustica.

Il primo sensore per il telerilevamento a ottenere un grande successo in meteorologia, a parte l'occhio umano e la macchina fotografica, è stato il radar a microonde (1950). Mentre la macchina fotografica si può considerare la capostipite dei telesensori passivi, vale a dire di quei sensori che utilizzano onde elettromagnetiche o acustiche emesse direttamente dall'ambiente in cui si opera, il radar è il primo esempio di una serie di strumenti in cui le radioonde vengono emesse artificialmente, interagiscono con il mezzo di cui si vogliono misurare le caratteristiche, e vengono quindi analizzate. L'attenuazione che tali onde subiscono nell'attraversare l'atmosfera esercita una limitazione, variabile con le condizioni di opacità, sulla distanza utile entro la quale lo strumento di telerilevamento può operare. Tale effetto è tanto più evidente quanto più la lunghezza d'onda della radiazione usata si avvicina alle dimensioni delle particelle che nell'atmosfera sono responsabili dell'attenuazione (da diffusione) del fascio in questione. Poiché in campo micrometeorologico sono molto spesso proprio tali particelle a essere l'oggetto di studio, è necessario operare, ove possibile, con potenze elevate e con rivelatori a elevata sensibilità.

La corsa alla realizzazione di strumenti sempre più potenti e raffinati che si verifica molto spesso con l'introduzione di una nuova tecnica di misurazione ha però qui trovato un ostacolo naturale nell'ambiente stesso in cui tali strumenti devono operare; poiché il loro impiego avviene spesso in presenza di forti escursioni termiche e inoltre richiede la loro collocazione su mezzi mobili è sempre necessario giungere a un compromesso nella scelta delle

Il lidar permette di determinare la posizione e le dimensioni di un pennacchio di fumo, la cui forma e sviluppo sottovento *(d)* sono strettamente legati ai parametri meteorologici che influenzano la diffusione atmosferica. D'altro canto, conoscendo la forma di un pennacchio, è possibile risalire all'andamento della direzione e velocità del vento e della temperatura relativi allo strato interessato dalla risalita del fumo. Si può notare nella figura che: la diminuzione della velocità del vento *(b)* è legata a un notevole aumento della dimensione verticale del pennacchio; il cambiamento di direzione del vento *(c)* si riflette sull'andamento della sezione orizzontale del pennacchio; la forma del pennacchio *(coning-fanning)* è tipica di un profilo di temperatura come quello in *a*.

Una interessante innovazione nelle tecniche antinquinamento è stata portata dalla introduzione del *remote sensing* attivo per la ricostruzione geometrico-strutturale dei pennacchi di fumo. L'accoppiamento lidar-spettrometro a correlazione permette di rilevare contemporaneamente sia la posizione, le dimensioni e i livelli di concentrazione di particolato solido (lidar), sia gli andamenti dello spessore ottico di gas inquinanti (spettrometro). Lo spessore ottico è dato dal prodotto della lunghezza del percorso sorgente-strumento per la concentrazione media di SO_2.

loro caratteristiche di potenza, ingombro, robustezza e affidabilità.

Come è stato già osservato in precedenza i parametri meteorologici fondamentali nel descrivere le caratteristiche fisiche dell'atmosfera sono le variazioni di temperatura, velocità e direzione del vento con la quota. I dati anemologici si ottengono dalla determinazione (mediante teodoliti, telemetri laser o radar) della traiettoria di un palloncino gonfiato con un gas più leggero dell'aria (elio o idrogeno) e lasciato salire liberamente nell'atmosfera.

I profili di temperatura si possono invece ottenere mediante l'uso di sonde radiotrasmittenti che vengono trascinate nell'atmosfera da un pallone sonda per trasmettere a terra i dati di pressione, temperatura e umidità dell'aria. La tecnica dei palloni sonda è ampiamente usata in meteorologia per conoscere la temperatura dell'aria fino ad altezze dell'ordine delle decine di chilometri. Poiché nel campo degli studi di diffusione atmosferica a scala locale si è interessati ad avere la maggiore risoluzione possibile entro uno strato limitato ai primi 500 o 600 metri di atmosfera, viene in genere impiegato un «pallone frenato». Il sistema è identico a quello precedente salvo il fatto che il pallone sonda non è lasciato libero di salire, ma è trattenuto da un cavo che ne regola quindi la salita e che permette di ricuperare, dopo ogni misurazione, la sonda. Quando le dimensioni del pallone lo consentono, questo sistema permette di collegare un certo numero di sonde al cavo di traino, ottenendo così profili più simultanei. Il sistema del pallone frenato ha una forte limitazione dovuta al fatto che, quando il vento supera un determinato valore di soglia (dipendente dalle caratteristiche aerodinamiche del pallone), non è più possibile innalzarlo.

Se ora passiamo dai sensori a misura diretta appena discussi (e che pur rappresentando un po' il «medioevo» della strumentazione meteorologica sono ancora oggi insuperati nel loro campo specifico) a quelli di telerilevamento troviamo, fra i più interessanti, i *lidar* e gli spettrometri a correlazione di maschera.

Il lidar (acronimo di *Light Detection And Ranging*) è un particolare tipo di radar che utilizza onde elettromagnetiche nel visibile, usando come sorgente di emissione un laser, per rivelare la presenza di aerosol o di particelle in sospensione nell'atmosfera. Il suo sviluppo (1965) si è reso possibile in seguito alla messa a punto dei laser a impulsi. La necessità della realizzazione di un tale strumento nasce dal fatto che il radar a onde centimetriche non permette, per le ben note leggi dello sparpagliamento della radiazione da parte dei centri diffusori, di «vedere» le particelle aventi dimensioni comprese fra 10^{-2} e 10^{2} micrometri. Queste ultime infatti retrodiffondono la radiazione elettromagnetica emessa dal radar all'incirca come le molecole dei gas atmosferici, confondendosi così nel cosiddetto «fondo molecolare». Per rilevare la presenza di aerosol è quindi necessario disporre di radiazioni elettromagnetiche le cui lunghezze d'onda siano dell'ordine di grandezza dei diametri delle particelle in esame e pertanto comprese nella regione dello spettro elettromagnetico del visibile o del vicino infrarosso (0,2-2 micrometri). Queste lunghezze d'onda, che sono quelle utilizzate dal lidar, permettono di distinguere lo sparpagliamento prodotto dalle molecole dei gas atmosferici (diffusione di Rayleigh) da quello generato dagli aerosol e dalle particelle inquinanti (diffusione di Mie).

Dall'analisi quantitativa e qualitativa degli echi lidar si rivela la presenza di particelle responsabili della retrodiffusione (aerosol, gocce d'acqua in sospensione, particelle di polveri) e se ne determina la posizione. È pertanto possibile utilizzare queste informazioni sia per rivelare particolari situazioni meteorologiche che ne favoriscono l'accumulo, sia per determinare la posizione e le caratteristiche geometriche e strutturali dei pennacchi di fumo o di nubi di inquinanti in funzione del tempo e della distanza dalla sorgente.

Nel primo caso lo strumento viene puntato in verticale e dall'analisi dell'eco si può calcolare sia lo spessore dello strato rimescolato che la quota a cui si trova la base di uno strato di inversione termica (presenza di un consistente addensamento di particelle deducibile da una discontinuità nell'eco). Questa tecnica, in parallelo con misurazioni di temperatura effettuate con l'uso di un pallone frenato, è stata utilizzata dal nostro gruppo in Val Padana.

Per la determinazione della posizione e delle dimensioni geometriche dei pennacchi di fumo o delle nubi di vapore emesse dalle torri di raffreddamento, il lidar deve invece essere posto in una posizione il più possibile perpendicolare all'asse del pennacchio per permetterne una buona prospezione. Tali misure hanno permesso, fra l'altro, di verificare la validità delle formule di innalzamento dei pennacchi riportate in letteratura ed, eventualmente, di modificarle.

La tecnica della spettrometria a correlazione di maschera, sviluppatasi in questi ultimi anni, permette di misurare a distanza le concentrazioni di alcuni inquinanti gassosi (in genere anidride solforosa e biossido d'azoto) ed è stata applicata con successo nel corso delle campagne sperimentali, sia perché fornisce misure integrate su una certa distanza (a differenza dei prelevatori puntuali che misurano solamente la concentrazione esistente in un determinato punto), sia perché consente una misura, effettuabile al suolo, della quantità di anidride solforosa (SO_2) o di biossido d'azoto (NO_2) presenti nell'atmosfera.

Tale tipo di spettrometria rappresenta una estensione all'atmosfera del metodo già adottato da tempo per misure in laboratorio, basato sul confronto tra lo spettro di assorbimento di un campione da analizzare e quello di un campione in cui il gas in esame è presente in concentrazione nota. Nella spettrometria a correlazione di maschera si fa uso di una sorgente emettente radiazioni luminose nella regione spettrale di assorbimento del gas (tale sorgente può essere sia una lampada sia il Sole o la radiazione diffusa dal cielo), e di un sistema di ricezione, in cui la radiazione in arrivo è confrontata, all'uscita di uno spettrometro, con una serie di fenditure fotoincise su una maschera (da cui il nome della suddetta tecnica) le quali riproducono lo spettro a bande del gas. Un sistema elettronico di elaborazione dei segnali all'uscita di un fotorivelatore consente infine di effettuare la misurazione voluta; si ottiene infatti un segnale significativo di correlazione solamente se il gas in esame è presente nel tratto di percorso atmosferico compreso tra la sorgente di radiazione e lo strumento. In questo modo si può risalire alla concentrazione media con cui l'inquinante esaminato compare nel tratto di cammino ottico che separa lo strumento dalla sorgente luminosa. L'uso di una sorgente di luce artificiale presenta l'indubbio vantaggio della precisa conoscenza delle caratteristiche del fascio luminoso emesso e della distanza tra il ricevitore e la sorgente. La sorgente artificiale è costituita generalmente da una lampada allo xeno con potenze variabili fra 500 e 2000 watt, modulata elettronicamente e dotata di un sistema ottico in configurazione di Cassegrain per concentrare maggiormente il fascio luminoso. Con tali lampade e in condizioni di particolare limpidezza atmosferica è stato possibile effettuare misurazioni su cammini ottici di 6-7 chilometri. Nella sua configurazione di strumento per il remote sensing passivo (uso del Sole o della luminosità diffusa del cielo quale sorgente di radiazione) lo spettrometro fornisce valutazioni quantitative sia della localizzazione che della concentrazione di alcuni inquinanti gassosi nell'atmosfera.

Grazie al suo ingombro ridotto, tale tipo di strumento è già da alcuni anni utilizzato dagli enti preposti al controllo delle emissioni industriali per una veloce localizzazione spaziale della presenza di SO_2 nell'atmosfera allo scopo di discriminare fra i contributi delle varie sorgenti presenti. In questo caso lo strumento, in puntamento verticale e configurazione passiva, viene montato su un automezzo che percorre a velocità costante un percorso di lunghezza nota misurando la quantità di gas presente sulla verticale. Il semplice passaggio al di sotto di un pennacchio a breve distanza dalla sorgente ha permesso di risalire in alcuni casi alla determinazione della quantità di inquinante emesso con una buona precisione (tutte le operazioni si svolgono in questi casi in tempo reale grazie all'aiuto di un piccolo calcolatore a bordo del mezzo).

La possibilità di impiegare spettrometri a correlazione di maschera quali elementi di reti di monitoraggio in sostituzione o in parallelo all'uso dei prelevatori puntuali è ancora oggetto di discussione, poiché nonostante il loro maggior raggio di azione richiedono un'assistenza pressoché continua e inoltre, fornendo una misura mediata sul percorso ottico, non possono rivelare i picchi di concentrazione localizzati in zone ristrette.

Canali ionici nella membrana delle cellule nervose

L'impulso nervoso è generato dal flusso di ioni sodio e potassio lungo i canali molecolari presenti nella membrana del nervo. Il funzionamento di questo tipo di canali è attualmente oggetto di studi approfonditi

di Richard D. Keynes

L'unità d'informazione fondamentale, trasmessa da una parte del sistema nervoso a un'altra, è un potenziale d'azione singolo, o impulso nervoso. Diversamente dai punti e dalle linee dell'alfabeto Morse, ogni impulso nervoso si presenta sempre con lo stesso aspetto: un'ampiezza di circa 100 millivolt e una durata fissa di un millisecondo circa. Il contenuto d'informazione dipende semplicemente dalla frequenza degli impulsi che una data fibra nervosa convoglia e dalle connessioni che questa fibra stabilisce a ogni estremità.

La linea di trasmissione per il potenziale d'azione è l'assone, o fibra nervosa, il lungo prolungamento cilindrico che emerge da ogni cellula nervosa periferica. In qualche misura, il meccanismo di trasmissione è legato alla somiglianza strutturale dell'assone con un cavo elettrico isolato, ma se la porzione centrale e il rivestimento di tale cavo avessero le resistenze elettriche tipiche di una fibra nervosa, il segnale si attenuerebbe rapidamente con la distanza. Il potenziale d'azione, invece, si automantiene e conserva un'ampiezza costante lungo tutto il suo percorso nella fibra. Tutto questo è in parte noto da tempo, ma i particolari molecolari sul modo in cui il processo si realizza sono ancora oggetto di studio.

Gli invertebrati che hanno bisogno di produrre una rapida risposta meccanica a una situazione di pericolo (ne sono esempi il colpo di coda di un'aragosta, la scomparsa nella sabbia di un granchio e la fuga, mediante propulsione a getto, di un calamaro), hanno sviluppato assoni giganti per assicurare una rapida conduzione degli impulsi che attivano la contrazione dei muscoli. Nel campo dell'elettrofisiologia, l'assone gigante per eccellenza è quello del calamaro: fornisce un'unica fibra nervosa facilmente sezionabile, con un diametro superiore a mezzo millimetro, in cui, per effettuare misurazioni elettriche, si possono inserire sistemi di elettrodi complessi e si può modificare la composizione ionica secondo la volontà dello sperimentatore.

L'evoluzione del sistema nervoso ha avuto un corso differente nei vertebrati, in cui la conduzione dell'impulso in molte fibre nervose è accelerata in misura considerevole dalla presenza di una guaina isolante costituita da un materiale lipidico, la mielina, la quale permette al suddetto impulso di saltare da un nodo di Ranvier (dove la mielina è interrotta per un breve tratto) all'altro. La mielinizzazione è una modificazione esclusivamente strutturale delle fibre e la membrana eccitabile dell'assone, nei punti in cui è a nudo, cioè in corrispondenza dei nodi di Ranvier di una fibra mielinica, si comporta esattamente come una fibra amielinica.

La presenza di una guaina mielinica conferisce un notevole vantaggio evolutivo ai vertebrati, perché permette una riduzione del diametro della fibra nervosa di 25 volte per raggiungere una data velocità di conduzione. Come conseguenza, però, le fibre mieliniche sono più difficili da isolare per dissezione e forniscono un materiale meno adatto delle fibre giganti amieliniche per studiare le proprietà delle membrane eccitabili.

Gli eventi fondamentali che stanno alla base della produzione di un potenziale d'azione sono stati chiariti 25 anni orsono nell'assone gigante di calamaro da Alan L. Hodgkin e da Andrew F. Huxley che lavoravano presso la Marine Biological Association a Plymouth in Inghilterra (per questa ricerca essi vinsero congiuntamente, nel 1963, il premio Nobel per la fisiologia o la medicina). Essi hanno dimostrato che l'eccitabilità elettrica della membrana nervosa dipende dal fatto che essa possiede un sistema di permeabilità ionica, che è sensibile all'intensità del campo elettrico e le consente di utilizzare l'energia immagazzinata nei gradienti di concentrazione ionica, stabiliti da una pompa ionica dipendente dall'energia del metabolismo.

Come nella maggior parte delle cellule viventi, l'interno dell'assone - l'assoplasma - contiene una notevole quantità di ioni potassio e pochi ioni sodio, mentre nei liquidi circolanti all'esterno dell'assone le concentrazioni relative di questi ioni sono invertite. Grazie a questa disuguale distribuzione di ioni e al fatto che la membrana allo stato di riposo è molto più permeabile agli ioni potassio che a quelli sodio, il nervo possiede un potenziale di riposo che varia dai 60 ai 70 millivolt, dato che l'interno è dotato di carica elettrica negativa rispetto all'esterno.

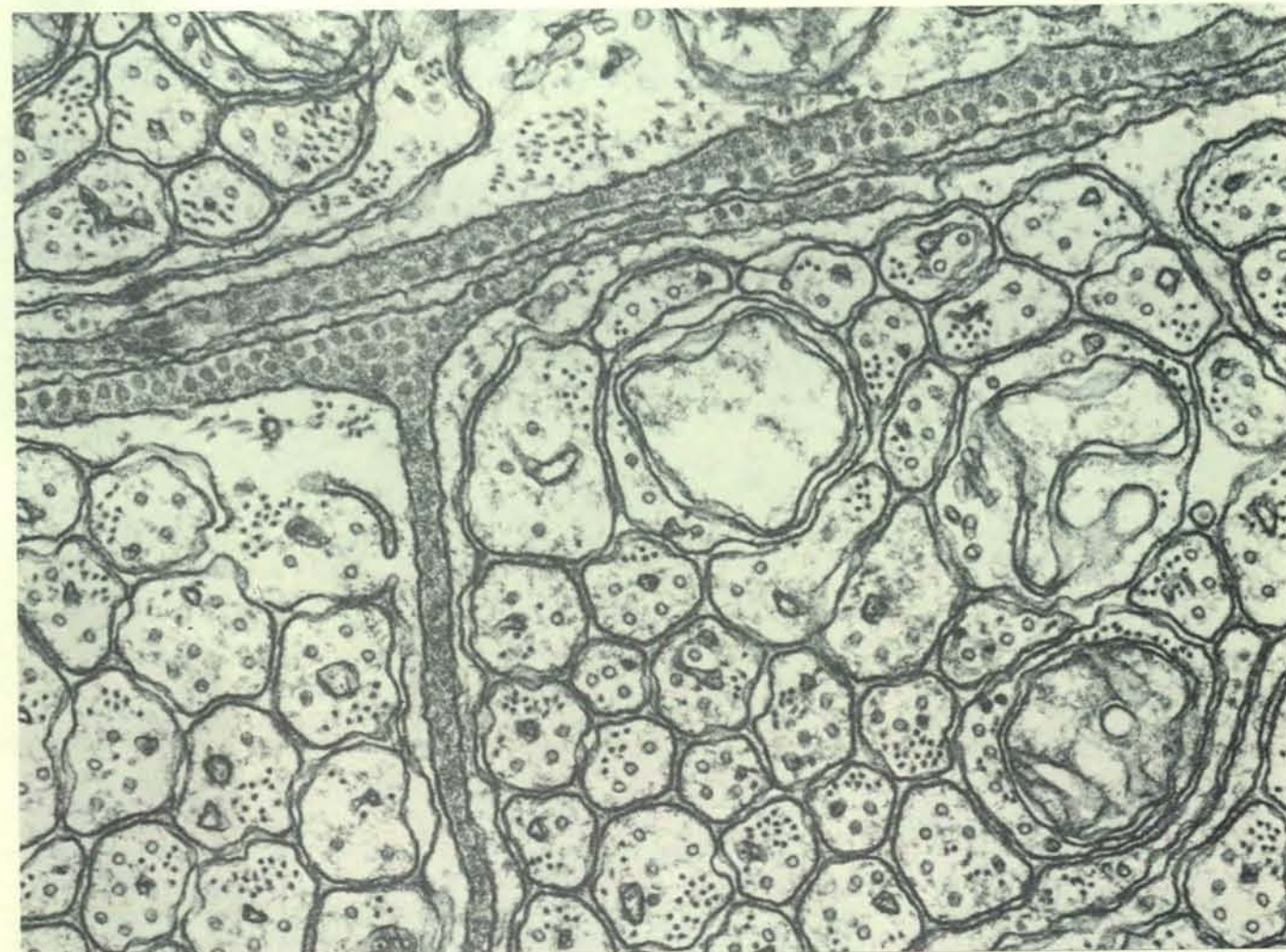
L'evento che scatena l'attività elettrica del nervo è una depolarizzazione della membrana, cioè una diminuzione della differenza di potenziale attraverso di essa, che può essere determinata o da un impulso eccitatorio da parte di un'altra cellula nervosa o da un impulso che proviene da ben più lontano, lungo la fibra. Questo fenomeno apre, nella membrana, dei canali che hanno una elevata selettività per gli ioni sodio e, di conseguenza, questi ultimi cominciano a fluire nella fibra, spinti da un elevato gradiente di potenziale elettrochimico diretto verso l'interno. L'ingresso di ioni dotati di carica positiva aumenta la depolarizzazione della membrana e pertanto vengono aperte ancora di più le barriere paragonabili a chiuse che controllano i canali del sodio per cui una volta dato avvio all'aumento di permeabilità per questo ione, questa cresce in maniera esplosiva fino a quando tutti i canali sono aperti. Come risultato, la differenza di potenziale transmembrana dapprima scende a zero e poi si inverte. Il flusso verso l'interno degli ioni sodio però rallenta man mano che ci si avvicina al potenziale d'equilibrio, cessando infine del tutto a causa dell'instaurarsi di un processo di inattivazione il quale chiude inesorabilmente i canali. Dopo un breve lasso di tempo, una seconda serie di canali, questa volta selettivi per il potassio, si dischiude e gli ioni potassio fluiscono verso l'esterno lungo il loro gradiente elettrochimico, fino a quando ricompare di nuovo l'originario potenziale di riposo.

Data la struttura paragonabile a un cavo dell'assone, l'onda di depolarizzazione si propaga lungo di esso a una velocità che - in un assone di calamaro di 600 micrometri di diametro o in un assone mielinico di rana di 12 micrometri - è di circa 20 metri al secondo. Nell'assone di calamaro, l'aumento di concentrazione del sodio interno durante un singolo impulso è molto piccolo e soltanto una milionesima parte del potassio interno va perduta; non si ha, pertanto, in seguito a questo fatto, alcun cambiamento misurabile nella velocità della pompa ionica. A lungo andare, invece, uno scambio di ioni che fornisca la fonte immediata di energia per la propagazione del potenziale d'azione dev'essere realizzato dalla pompa ionica, la quale viene alimentata dal metabolismo cellulare.

Per riuscire a spiegare questa sequenza di eventi, Hodgkin e Huxley hanno postulato l'esistenza di serie di canali paralleli, selettivi o per gli ioni sodio o per quelli potassio e controllati dal campo elettrico stabilitosi nello spessore della membrana. Essi hanno quindi proceduto all'elaborazione di un metodo che permettesse di realizzare studi quantitativi sulle proprietà dei canali per il sodio e per il potassio. Lo scopo della tecnica, da loro chiamata *voltage clamp*, era quello di rendere possibile una modificazione istantanea del potenziale di membrana verso un nuovo livello prestabilito, misurando contemporaneamente la corrente necessaria per mantenere il potenziale a quel livello. Adottando questa tecnica essi seguivano l'idea sviluppata originariamente da K. S. Cole e G. Marmont.

Tale tecnica comportava l'inserimento nell'assone di due elettrodi, uno per misurare il potenziale interno e l'altro collegato a un elettrodo esterno in modo da creare un flusso di corrente attraverso la membrana. Questi elettrodi venivano inseriti nel circuito a retroazione di un amplificatore in modo che il potenziale registrato potesse essere confrontato con la tensione stabilita dallo sperimentatore; nell'assone veniva così fatta fluire una quantità di corrente sufficiente per portare al livello prestabilito qualsiasi differenza di potenziale tra l'elettrodo interno e quello di controllo.

Il potenziale di membrana poteva così essere mantenuto a un livello, o a una sequenza di livelli, ambedue prestabiliti. La dipendenza della corrente ionica dal tempo e dalla tensione era misurabile in varie condizioni sperimentali. Da quando venne introdotta, questa tecnica ha subito diversi adattamenti per poter essere applicata alle fibre nervose mieliniche e a molti altri tipi di tessuti eccitabili; in effetti, è lo strumento principale a cui si fa ricorso nella maggioranza degli studi sulle proprietà delle membrane eccitabili.

Le registrazioni ottenute da Hodgkin e Huxley con tale tecnica hanno messo in luce che i sistemi sodio e potassio si comportavano in un modo nettamente diverso quando la membrana era depolarizzata. La permeabilità per il sodio saliva rapidamente fino a un massimo e quindi scendeva fino a zero, allorché si instaurava il fenomeno di inattivazione; la permeabilità per il potassio, invece, cresceva con un apprezzabile periodo di latenza, ma non si azzerava, indipendentemente dal periodo di depolarizzazione. La durata dei cambiamenti a carico della permeabilità è stata analizzata nei particolari e i risultati sono stati introdotti in una famosa serie di equazioni che descrivono in termini quantitativi le proprietà dei due

Microfotografia elettronica di fasci di assoni del nervo olfattivo di luccio, in sezione trasversale. Ogni assone è un tubo delimitato da membrana e ripieno di un liquido viscoso, l'assoplasma. La membrana presenta canali molecolari distinti per il passaggio degli ioni sodio e potassio, il cui flusso attraverso di essa genera l'impulso nervoso. Nell'assone di luccio, i canali per ambedue gli elementi sono scarsi: la loro densità è di soli pochi canali per micrometro quadrato. La microfotografia, che ingrandisce le fibre 80 000 volte, è di A. von Muralt dell'Università di Berna.

tipi di canali ionici. Ciò ha permesso a Hodgkin e a Huxley di raggiungere il loro obiettivo primario, che era quello di mettere alla prova la correttezza della loro analisi prevedendo, sulla base delle loro equazioni, la conformazione precisa del potenziale d'azione convogliato dalla fibra e anche caratteristiche come la velocità di conduzione e la quantità di ioni sodio e potassio scambiati. Il successo ottenuto con tali calcoli fu davvero trionfale.

Questa classica ricerca costituì un'importante progresso, che lasciò, tuttavia, senza risposta due interrogativi fondamentali. Innanzitutto, come fanno i canali a discriminare in maniera così marcata tra ioni dotati tutti di una singola carica positiva, come sono gli ioni di sodio, di potassio e di altri metalli alcalini? Gli ioni litio penetrano nei canali del sodio all'incirca con la stessa facilità degli ioni sodio, mentre gli ioni potassio vi riescono con una velocità che è solo la dodicesima parte di quella degli ioni sodio; per il rubidio e per il cesio, tale frazione è troppo piccola per poter essere misurata. In secondo luogo, in che modo l'apertura e la chiusura dei canali dipende in misura così stretta e con una coincidenza così precisa dal campo elettrico che si instaura attraverso la membrana?

Il meccanismo che è alla base della specificità dei canali è stato studiato da Bertil Hille dell'Università di Washington in una elegante serie di esperimenti, eseguiti mediante la tecnica messa a punto da Hodgkin e Huxley, sulle fibre nervose mieliniche di rana. Hille ha esaminato la permeabilità del nodo di Ranvier a un'ampia varietà di piccoli ioni organici positivi e ha trovato che soltanto molecole misuranti meno di 0,3 per 0,5 nanometri potevano passare attraverso i canali ionici. In parte, dunque, la selezione dipendeva da una buona concordanza tra le dimensioni dello ione che penetrava e quelle del canale. Tuttavia, anche tra ioni positivi delle dimensioni appropriate esistevano sorprendenti disparità di permeabilità: per esempio l'idrossilammina ($OH\text{-}NH_3^+$) e l'idrazina ($NH_2\text{-}NH_3^+$) passavano senza difficoltà lungo il canale, mentre la metilammina ($CH_3\text{-}NH_3^+$) non passava. Per spiegare questa discrepanza, Hille ha avanzato l'ipotesi che i canali del sodio fossero tappezzati, nel punto più stretto, da atomi di ossigeno, con carica negativa. Gli ioni positivi contenenti gruppi ossidrilici (OH) o amminici (NH_2) potevano scivolare lungo il canale formando legami idrogeno con gli atomi di ossigeno, mentre i gruppi metilici (CH_3), che non formavano questi legami, erano troppo grossi per poter passare.

Nel corso di misurazioni della corrente ionica, effettuate quando il sodio all'esterno era stato sostituito da altri ioni, capaci o no di passare attraverso i suoi canali, Hille è riuscito a dimostrare che gli ioni sodio si legavano in qualche punto del canale a siti fissi, dotati di carica negativa; per questo il loro comportamento è noto come progressione «in fila indiana». Il fenomeno, per cui gli ioni passano lungo il canale uno dietro l'altro, fu osservato per la prima volta da Hodgkin e da me, alcuni anni fa, in esperimenti effettuati con traccianti radioattivi, sui canali del potassio negli assoni giganti di calamaro. Hille ha suggerito che esistono quattro barriere energetiche nei canali del sodio e che solo uno ione alla volta viene fatto passare all'interno.

La più elevata di queste barriere energetiche costituisce il filtro di selettività. La relativa facilità con cui gli ioni sodio idratati riescono a superarlo dipende dalla loro capacità di subire una graduale rimozione delle molecole d'acqua stabilizzatrici, interagendo con lo ione COO^- presente nella parete del canale. Gli ioni potassio, che sono un poco più grossi, non possono perdere con altrettanta facilità il loro involucro di molecole d'acqua e, pertanto, si trovano di fronte una barriera energetica molto più alta. Si va così stabilendo una convincente rappresentazione della disposizione delle molecole in quella porzione di canale del sodio che è responsabile della sua selettività.

Buona parte dei recenti progressi verso la comprensione del sistema di barriere dei canali del sodio sono stati compiuti con il sussidio di un importante veleno nervino, la tetrodotossina. Questa molecola si lega all'imboccatura del canale del sodio, rivolta verso l'esterno, con un'affinità che la rende una delle sostanze più tossiche che si conoscano, e blocca il passaggio degli ioni sodio. Si trova in notevoli quantità nelle ovaie, nel fegato, nella cute e nell'intestino (in minor misura anche nei muscoli) dei pesci della famiglia tetraodontidi, tra cui la specie *Spheroides porphyreus* è nota ai giapponesi con il nome di *fugu*. Il fugu, in Giappone, è ritenuto una vera e propria prelibatezza, ma prima di mangiarlo occorre assicurarsi che il cuoco sia perfettamente al corrente del procedimento indispensabile per scartare dal pesce tutte le parti tossiche. La tetrodotossina è anche presente in due animali che non hanno alcun rapporto di parentela con i tetraodontidi, il tritone californiano *Taricha torosa* e il polpo australiano *Hapalochlaena maculosa;* inoltre un analogo, che differisce dalla tetrodotossina solo per un gruppo laterale, è stato isolato dalla rana centroamericana *Atelopus chiriquensis.*

Due tipi di canali ionici attraversano la membrana dell'assone. Una «pompa» dipendente dall'energia metabolica, alimentata dall'idrolisi dell'ATP, trasporta controcorrente gli ioni sodio *(cerchietti in colore)* e gli ioni potassio *(cerchietti in nero)* e stabilisce, attraverso la membrana, gradienti di concentrazione per questi ioni. Il secondo tipo di canale permette agli ioni di fluire nel senso del gradiente in risposta a cambiamenti di potenziale elettrico nello spessore della membrana. In questo schema, vengono rappresentati i due canali del sodio e del potassio, nella condizione di riposo, cioè con barriere tenute in funzione dal potenziale di membrana. Quando il potenziale di riposo si riduce, i canali si aprono (più rapidamente quelli del sodio, più lentamente quelli del potassio) e danno origine a un impulso di corrente, che si propaga lungo l'assone. In questo articolo ci occupiamo esclusivamente di quei canali che permettono movimenti di ioni nel senso del gradiente.

L'impulso nervoso può essere rappresentato da variazioni del potenziale della membrana assonica *(a)* o della conduttanza della stessa agli ioni sodio e potassio *(b)*, ambedue espressi su una scala del tempo in millisecondi. Durante l'impulso, l'interno della membrana diventa per breve tempo positivo rispetto all'esterno. Quando l'impulso è cessato, il potenziale di riposo viene ristabilito.

Un altro veleno paralizzante, di tossicità altrettanto elevata e che agisce nello stesso modo della tetrodotossina, malgrado la sua struttura diversa, è la saxitossina. Questa molecola viene elaborata da una specie del genere *Gonyaulax*, che fa parte del plancton marino e, in condizioni favorevoli, può moltiplicarsi così rapidamente da raggiungere una densità di molti milioni di individui per litro, provocando il fenomeno detto «marea rossa», il quale trasforma il colore della superficie degli oceani in un rosso ruggine che, di notte, diviene luminescente. I crostacei che si alimentano nelle acque che presentano questa fioritura di *Gonyaulax* concentrano la tossina nei loro tessuti e, pertanto, diventano velenosi per l'uomo: basta ingerire una quantità di solo 1 milligrammo di saxitossina perché si verifichi una paralisi fatale. La comparsa indipendente di molecole insolite, come quelle della tetrodotossina e della saxitossina, in specie assai diverse, grazie a processi casuali di selezione naturale, costituisce un valido esempio di evoluzione convergente.

La tetrodotossina è una molecola organica complessa, costituita da parecchi anelli congiunti, a una estremità dei quali è presente un gruppo guanidinico, dotato di carica positiva. È probabile, a quanto sembra, che tale gruppo simuli lo ione sodio, dotato di uguale carica elettrica, per cui la molecola può oltrepassare l'imboccatura del canale del sodio, dove poi si blocca, tappando il canale, come fa un uomo grasso che non riesce a passare attraverso un oblò.

Per spiegare perché la tetrodotossina agisce a una concentrazione così straordinariamente bassa, è tuttavia necessario supporre che il resto della molecola si conformi in maniera perfetta alla superficie esterna del canale. Per spiegare poi il legame altrettanto stretto della saxitossina con i suoi due gruppi guanidinici e anelli diversi, è necessario supporre che la molecola occupi un recettore vicino, ma non identico. Certamente il recettore per la tossina è localizzato soltanto all'esterno del canale, dato che una soluzione anche concentrata di tetrodotossina non ha alcun effetto bloccante quando viene introdotta all'interno di una assone. Sia la tetrodotossina, sia la saxitossina sono specifiche per i canali del sodio; né l'una né l'altra hanno il più lieve effetto sulla conduttanza dei canali del potassio.

Quando un assone si trova in presenza di tetrodotossina o di saxitossina, solo una molecola di queste sostanze si lega a ogni canale; pertanto esse servono mirabilmente a contare il numero di canali presenti nella membrana. Questo esperimento è stato realizzato per la prima volta da John Moore, Toshio Narahashi e Trevor Shaw, che lavoravano alla Duke University. Essi hanno progettato un ingegnoso metodo di analisi biologica per la misurazione indiretta del legame della tetrodotossina ai piccoli assoni del nervo di una zampa di aragosta e hanno trovato che, in questo tipo di nervo, i canali del sodio erano piuttosto pochi e distanziati, essendovene solo tredici per micrometro quadrato di membrana.

Quest'osservazione è stata successivamente confermata e estesa ad altri tipi di assoni da J. Murdoch Ritchie e collaboratori della Yale University School of Medicine. Essi hanno utilizzato tetrodotossina e saxitossina che incorporavano atomi di trizio, l'isotopo radioattivo dell'idrogeno, per cui la quantità di tossina legata alla membrana poteva essere misurata direttamente. I loro risultati hanno dimostrato che gli assoni più piccoli, quelli dei nervi olfattivi dei pesci, hanno il più basso numero di canali del sodio per micrometro quadrato di membrana (solo tre circa), mentre le fibre, un poco più grosse, del nervo vago del coniglio ne hanno 25 per micrometro quadrato. Lavorando presso la Marine Biological Association di Plymouth, Hans Meves e Rock Levinson hanno trovato che l'assone gigante di calamaro ha una densità molto più elevata di canali del sodio: fino a 500 per micrometro quadrato. Questa cifra più elevata concorda con altre misure, in particolare con quelle ottenute per il flusso del sodio. Essa non è molto diversa dal numero di canali calcolato da Hodgkin per spiegare la massima velocità di conduzione dell'impulso che si propaga in un assone del diametro di 500 micrometri.

Un secondo pregio della tetrodotossina dal punto di vista degli elettrofisiologi consiste nel fatto che permette di agire direttamente sul meccanismo che blocca il passaggio di ioni. Come hanno sottolineato, nel 1952, Hodgkin e Huxley, la forte dipendenza della permeabilità ionica della membrana dalla differenza di potenziale che si instaura nel suo spessore significa che i canali del sodio e del potassio devono essere controllati dai movimenti di «chiuse» che si comportano come se avessero una grossa carica elettrica. Al normale potenziale di riposo di

circa −65 millivolt, la maggioranza delle particelle cariche con funzione di barriera si trovano nella configurazione di riposo, o di chiusura, e i canali sono sbarrati. Come risposta a una variazione del potenziale di membrana verso un valore più positivo, cioè verso una depolarizzazione, tali particelle cominciano a modificare la loro configurazione verso la forma «aperta» e gli ioni possono così passare attraverso i canali.

Il funzionamento del sistema sopradescritto deve necessariamente comportare la migrazione di particelle cariche in un campo elettrico. Così, si potrebbe pensare che il cambiamento di configurazione a cui si è accennato preceda l'apertura dei canali, in modo da provocare un piccolo spostamento di cariche attraverso la membrana. In altre parole, ci si aspetterebbe di vedere un flusso transitorio dovuto a una corrente di spostamento che precede il passaggio della corrente ionica. Hodgkin e Huxley avevano previsto l'esistenza di questa corrente, ma non furono in grado di misurarla per la semplice ragione che, a quell'epoca, non esisteva alcun mezzo per impedire che l'ampio flusso di ioni attraverso i canali del sodio l'oscurasse completamente. Nel breve periodo in cui un singolo canale del sodio, durante il passaggio di un impulso nervoso, rimane aperto, circa 100 ioni sodio passano in esso; durante un lungo impulso, qual è quello che si somministra con la tecnica del *voltage clamp*, potrebbero passare in esso fino a 500 ioni. L'apertura della barriera, d'altra parte, comporta il trasferimento di non più di quattro cariche elettriche all'interno della membrana.

Il potenziale che determina la chiusura dei canali è molto più basso di quello che consente il flusso degli ioni e, pertanto, non può facilmente essere riconosciuto se prima non si ostruiscono accuratamente i canali ionici. Avendo a disposizione la tetrodotossina, si sono potute realizzare condizioni più favorevoli alla sua registrazione, grazie anche alla fortunata circostanza per cui, quando la tossina si lega all'imboccatura del canale del sodio, bloccandolo, cioè non impedisce né l'apertura e la chiusura della barriera stessa né il flusso della corrente.

Allora, se il flusso di sodio in un assone di calamaro viene abolito applicando una dose piuttosto grande di tetrodotossina (e sostituendo il sodio nel mezzo che bagna l'assone con tampone Tris) e, nel contempo, i canali per il potassio sono bloccati dall'interno perfondendo l'assone con ioni cesio o tetraetilammonio, la corrente di sbarramento costituisce una frazione importante della corrente di spostamento, registrata quando il potenziale a cui la membrana viene bloccata è improvvisamente alterato. Con questo metodo, correnti di sbarramento per il sodio sono state osservate per la prima volta nel calamaro da Clay Armstrong e Pancho Bezanilla del Marine Biological Laboratory of Woods Hole, nella stagione 1972-73, e poco dopo da Eduardo Rojaz e da me a Plymouth. Come in altri casi, si vide che osservazioni come queste, effettuate sull'assone gigante di calamaro, erano facilmente estendibili ad altri tipi di neuroni e, pertanto, studi sulla corrente di sbarramento in corrispondenza dei nodi di Ranvier di un nervo di rana furono effettuati con successo in varie sedi, tra cui l'Università di Amburgo, a opera di Wolfgang Nonner, Berthold Neumcke e Robert Stämpfli.

Anche se, con il sussidio della tetrodotossina, la corrente di sbarramento può essere distinta dalla corrente ionica senza eccessiva difficoltà, resta il problema di distinguere la corrente di spostamento, che deriva da un caricarsi e scaricarsi dell'elevata capacità statica della membrana, dalla corrente di spostamento formatasi a seguito dei movimenti delle particelle che costituiscono la barriera. Si può ragionevolmente ammettere che, per ottenere un reale transitorio della capacità della membrana, la quantità totale delle cariche che passano attraverso di essa a seguito dell'applicazione di un impulso per ottenere un *voltage clamp*, varia in proporzione diretta alla dimensione del salto di potenziale. Pertanto le oscillazioni transitorie della capacità per impulsi di uguali dimensioni, ma di direzione opposta, saranno sempre esattamente simmetriche.

D'altra parte, le particelle che operano lo sbarramento avranno all'inizio una distribuzione disuguale tra le loro due configurazioni e, se il loro numero totale è fisso, lo spostamento di cariche, quando cambiano stato, mostrerà una saturazione se gli impulsi saranno sufficientemente ampi. Di conseguenza, l'ampiezza della corrente di sbarramento sarà asimmetrica rispetto al potenziale. Partendo da un potenziale negativo, in cui le particelle che fanno da barriera sono tutte nella configurazione «chiusa», la registrazione della corrente di spostamento per un impulso iperpolarizzante, impartito per portare il potenziale a un livello prestabilito più negativo, consisterà in una pura e semplice oscillazione transitoria della capacità della membrana, dato che le barriere non possono chiudersi ulteriormente. La corrente di spostamento per un impulso depolarizzante di medesima ampiezza consisterà, invece, in un uguale transitorio di capacità di segno opposto, sommato alla corrente di sbarramento, generata dall'apertura delle barriere.

Il metodo di registrazione della corrente di sbarramento, adottato all'inizio da Armstrong e Bezanilla a Woods Hole e da Meves e me a Plymouth, consisteva semplicemente nel sommare le registrazioni delle correnti dovute a impulsi uguali e opposti somministrati per portare il potenziale a un livello prestabilito in assoni con canali ionici chiusi, cancellando così automaticamente l'oscillazione transitoria della capacità della membrana. Da allora, la tecnica è stata migliorata, per certi aspetti, ricorrendo a un calcolatore ma il principio rimane lo stesso. La corrente di sbarramento sale fino a un valore massimo subito dopo l'applicazione della differenza di potenziale, quindi scende a zero con un andamento esponenziale a mano a mano che le particelle che costituiscono la barriera assumono la loro nuova configurazione. Con il crescere dell'ampiezza dell'impulso, aumenta la carica elettrica spostata, mentre diminuisce il tempo di riposo.

Una conclusione che si può trarre considerando le proprietà allo stato di equilibrio delle particelle che fungono da barriera è che, come Hodgkin e Huxley avevano suggerito in precedenza, sulla base di prove cinetiche, il meccanismo di sbarramento comporta la cooperazione di tre (o probabilmente quattro) particelle cariche per ogni canale. Le singole particelle si comportano come se avessero un'unica carica elettrica, ma potrebbe benissimo darsi che, in realtà, esse abbiano un certo numero di gruppi laterali carichi, che si spostano solo di poco nel campo elettrico. Non è possibile stabilire, in base ai dati disponibili, la distribuzione dei gruppi laterali all'interno della particella. Tuttavia, confrontando la velocità con cui le particelle che formano una barriera raggiungono la loro nuova configurazione con quella del passaggio degli ioni attraverso i canali, è evidente che la cinetica della corrente di sbarramento non si concilia direttamente con le equazioni di Hodgkin-Huxley. Perché i dati possano concordare, si deve supporre che, tra le configurazioni «chiusa» e «aperta», vi sia una configurazione «attivata» e che, dopo che le particelle l'hanno assunta, segua un periodo ben circoscritto in cui le particelle interagiscono, cooperando tra loro, prima che i canali si aprano. In questo periodo, la corrente di sbarramento non esiste più.

L'altra importante conclusione riguarda il meccanismo di inattivazione dei canali del sodio: è questo il processo che chiude di nuovo i canali all'incirca un millisecondo dopo che essi sono stati aperti per la prima volta, e ciò avviene anche se la membrana rimane depolarizzata. Non si ritiene più, ormai, che, oltre alle particelle cariche che aprono i canali, esista una serie separata di particelle che li bloccano con un opportuno sfasamento di tempo. Il motivo è che queste particelle che operano il blocco dovrebbero anch'esse essere cariche, mentre non si riesce a registrare alcuna corrente asimmetrica di spostamento, corrispondente al loro movimento. Armstrong e Bezanilla hanno proposto che l'inattivazione sottintenda una ulteriore ridistribuzione delle particelle che fungono da barriera e che essa faccia lentamente chiudere i canali di conduzione. Si ritiene che questo cambiamento finale nella configurazione sia accompagnato da uno scarso movimento di cariche, o addirittura da nessun movimento, e che sia, pertanto, elettricamente muto. Esso lascia chiaramente le particelle che fungono da barriera in una condizione di parziale immobilità, con il risultato che, al momento della ripolarizzazione della membrana, il loro ritorno allo stato di riposo è troppo lento perché possa essere prontamente registrato.

Lo studio della corrente di sbarramento per il sodio aveva dunque condotto al concetto secondo cui una porzione del canale per questo elemento, sensibile alla differenza di potenziale, sarebbe un sistema in cui molteplici subunità passano attraverso svariate configurazioni. Si può costruire un modello matematico che tenga conto dei principali risultati che ho descritto, ma esso avrà probabilmente un limitato interesse per gli sperimentatori in quanto, come tutti i modelli che si riferiscono a più stati, include troppi parametri arbitrari per fornire al problema un'unica soluzione. Ciononostante, non si può negare che, dal punto di vista operativo, i canali possono essere: 1) in stato di riposo; 2) in fase di conduzione; 3) in uno stato di inattivazione o di refrattarietà. Di conseguenza, piaccia o no, il sistema esiste necessariamente in almeno tre stati diversi. Altri stati intermedi possono pure rendersi necessari se si devono opportunamente fissare nel tempo le transizioni tra gli stati principali. C'è da sperare che, da ulteriori e più accurate misurazioni della corrente di sbarramento, potranno emergere informazioni su una o più di queste transizioni ed esse le caratterizzeranno indipendentemente dai dati derivati dalla cinetica globale del sistema.

La costruzione di un modello molecolare plausibile per i canali per il sodio è ancor più lontana oggi dalle nostre possibilità. È chiaro che il filtro di selettività è localizzato vicino all'estremità del canale rivolta all'esterno e sembra che, in corrispondenza del suo punto più stretto, la struttura del canale sia rigida e insensibile alle modificazioni del potenziale di membrana. La difficoltà sta nel rappresentare la porzione centrale del canale, con la sua barriera di potenziale, dato che non esistono precedenti che permettano anche solo di fare una supposizione sul modo in cui una modificazione del campo elettrico cui è sottoposta una grossa molecola possa aprire in essa un canale acquoso, per non dire di una molecola la cui apertura e chiusura richiede una cinetica così complessa. Pertanto qualsiasi rappresentazione dell'attività interna dei canali per il sodio non può che essere speculativa.

LITIO

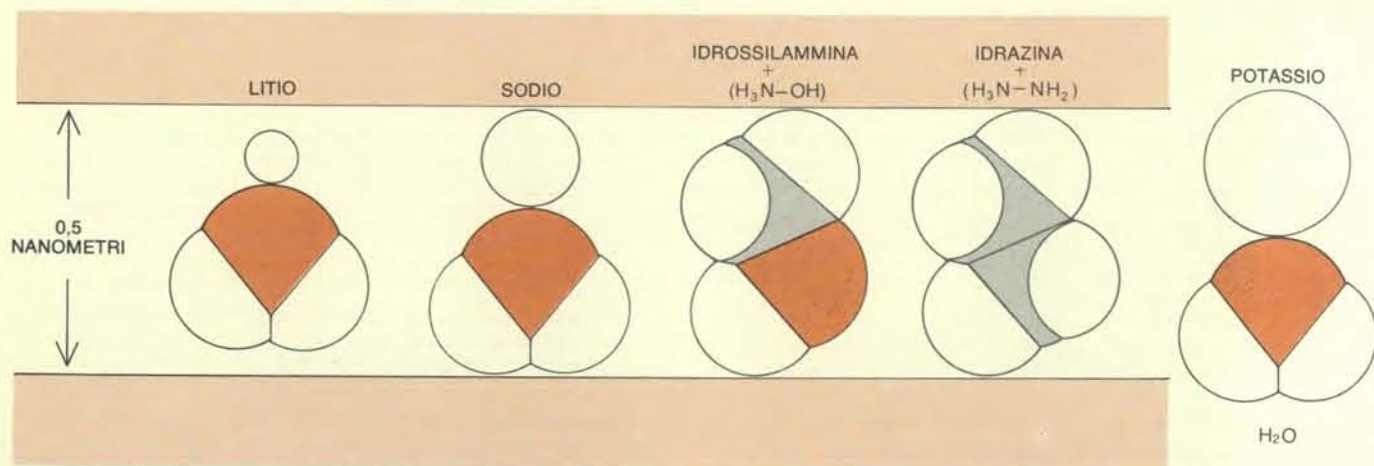
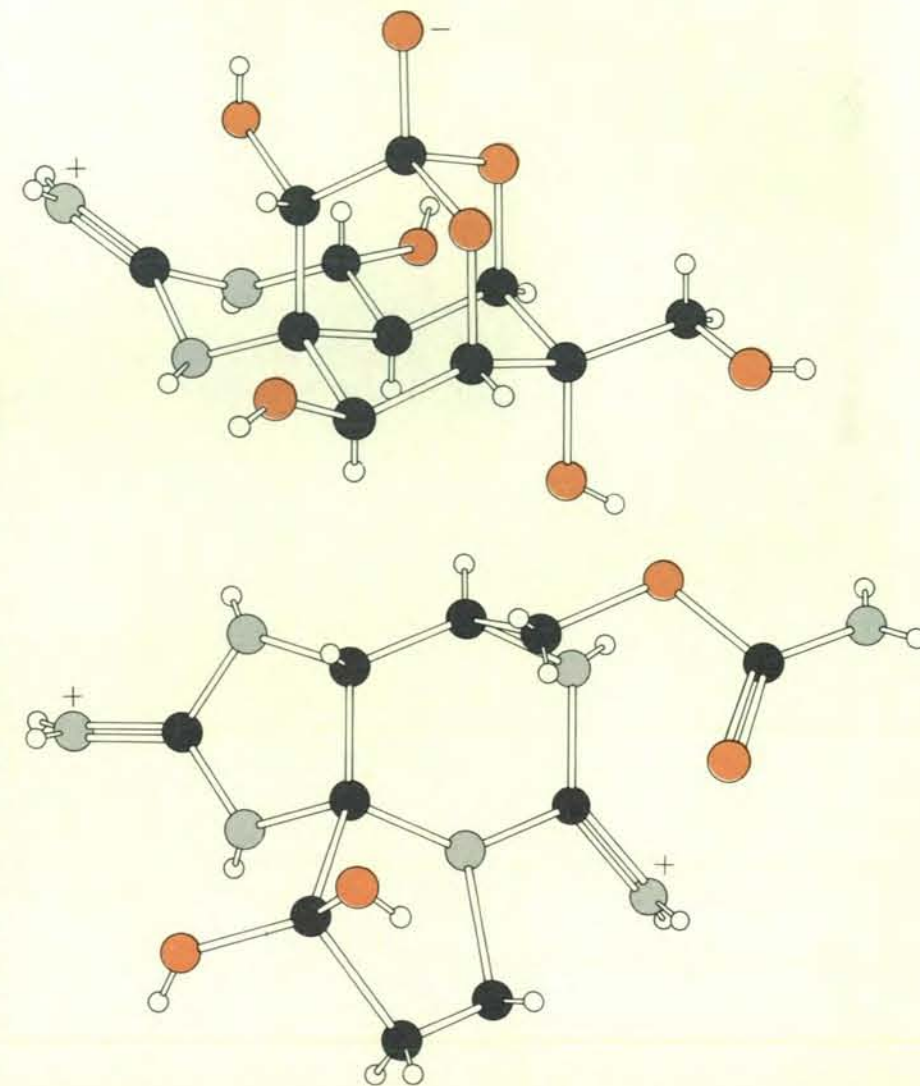
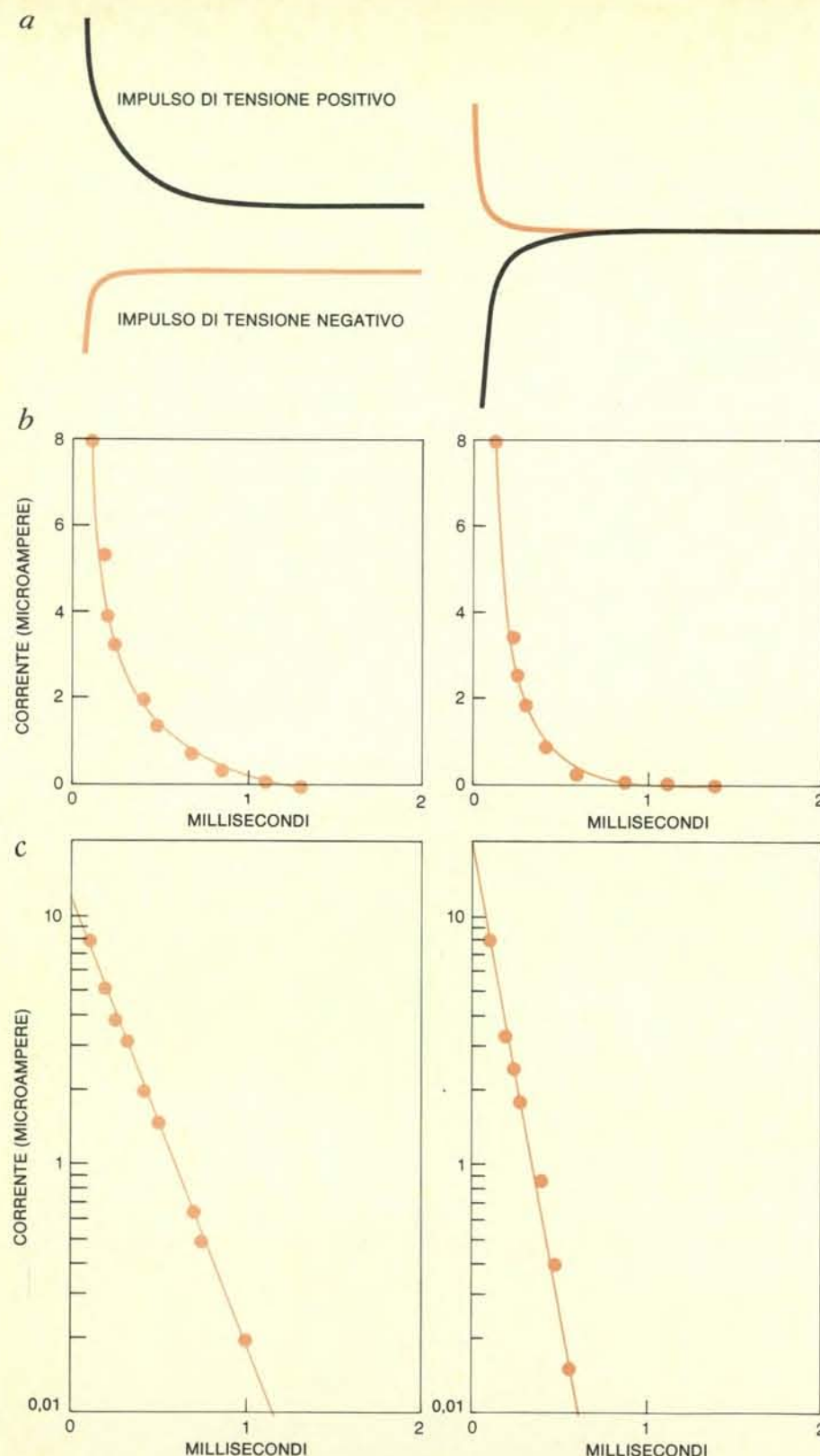
SODIO

IDROSSILAMMINA
($H_3\overset{+}{N}$–OH)

IDRAZINA
($H_3\overset{+}{N}$–NH_2)

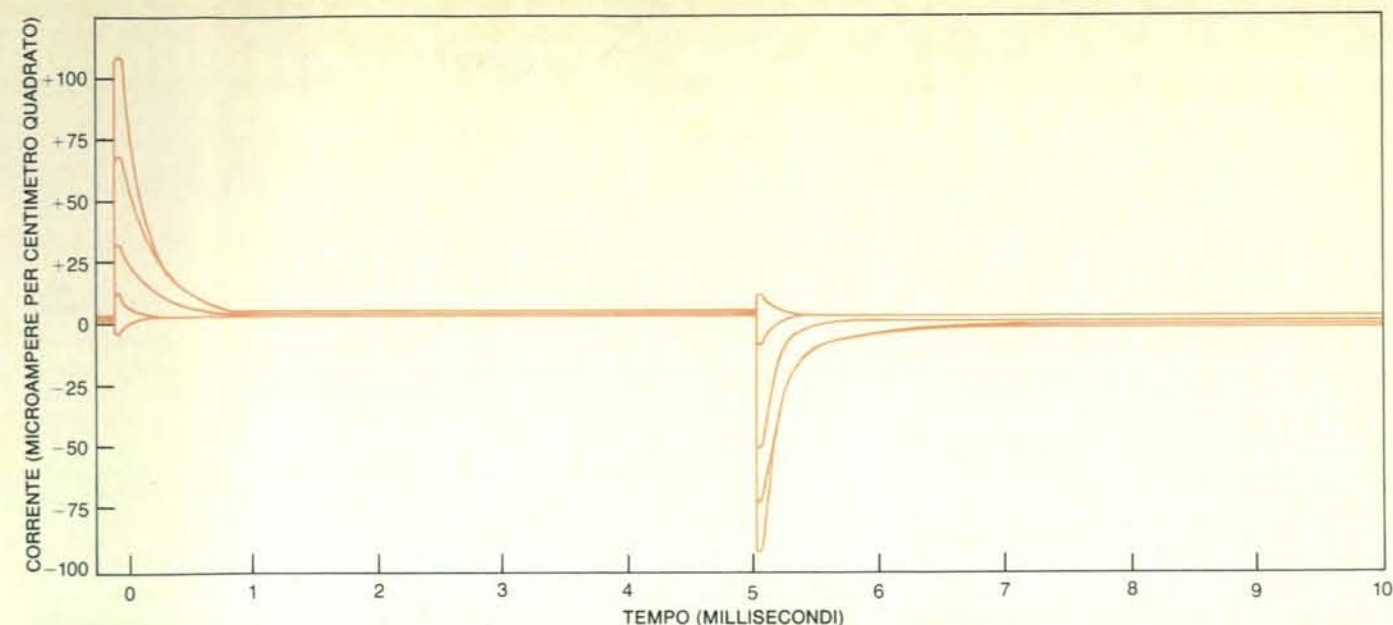
POTASSIO

0,5
NANOMETRI

H_2O

La selettività ionica del canale del sodio è dovuta alle dimensioni dello ione di questo elemento e alle sue proprietà elettrostatiche. Come mostra schematicamente il disegno, gli ioni litio o sodio, parzialmente idratati, e le piccole molecole, dotate di carica positiva, dell'idrossilammina e dell'idrazina hanno le dimensioni giuste per passare attraverso il canale del sodio, che risulta avere una sezione trasversale di 0,3×0,5 nanometri nel suo punto più stretto. Uno ione potassio idratato, invece, è troppo grosso per riuscire a passare attraverso quel canale.

Modelli della tetrodotossina *(in alto)* e della saxitossina *(in basso)*, due veleni nervini che si legano in maniera tenace e specifica ai canali del sodio. Essi rivelano alcune caratteristiche strutturali in comune: in ambedue un gruppo dotato di carica positiva imita lo ione sodio, pure con carica positiva, entrando nel suo canale, ma la struttura policiclica massiccia impedisce alla tossina di passare. Per questa loro proprietà, entrambe le sostanze si sono dimostrate utili per lo studio di questi canali.

La corrente di sbarramento è associata al movimento di particelle o subunità dotate di carica elettrica e funzionanti come barriere all'interno dei canali del sodio. Queste subunità aprono o chiudono i canali in risposta a variazioni del potenziale di membrana. La sopracitata corrente può essere riconosciuta soltanto dopo aver bloccato la corrente del sodio con la tetrodotossina e quella del potassio con il cesio, e sbilanciando il transitorio simmetrico della capacità della membrana. In queste condizioni, impulsi di corrente, uguali e di carica opposta danno origine attraverso alla membrana a flussi di corrente leggermente diversi: questa differenza costituisce la corrente di sbarramento. I tracciati in *a)* sono registrazioni di un'unica onda della corrente di membrana per impulsi di ± 120 millivolt. La differenza tra loro è rappresentata graficamente in *b)* in scala lineare e in *c)* in scala logaritmica. Uno spostamento del potenziale di membrana verso valori più elevati (ma non verso valori negativi) provoca una breve corrente di sbarramento verso l'esterno, che coincide con l'apertura dei canali del sodio. Tale corrente cade a zero con un andamento esponenziale a mano a mano che le subunità di sbarramento raggiungono la loro nuova configurazione.

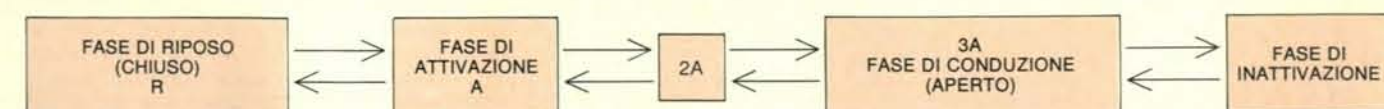
La mia descrizione ha sin qui riguardato quasi per intero i canali del sodio, ma ciò non significa che quelli del potassio siano meno importanti. Tuttavia, per due motivi non si hanno su di essi altrettante notizie. Innanzitutto, non esiste per i canali del potassio alcun agente bloccante che abbia qualcosa di analogo alla specificità che la tetrodotossina ha per i canali del sodio. Anche se gli ioni tetrametilammonio, dotati di carica positiva, possono penetrare nei canali del potassio dall'interno e tapparli bene, è necessaria una concentrazione relativamente elevata perché essi possano produrre un blocco completo. Non è così possibile ricorrere all'espediente di applicare un tracciante radioattivo alle molecole bloccanti per contare i canali del potassio. In secondo luogo, dato che i canali del potassio si aprono con ritardo e con una velocità circa 10 volte inferiore rispetto ai canali del sodio, la corrente di sbarramento per il potassio è molto più difficile da rilevare.

Per un aspetto, tuttavia, la lentezza con cui si aprono i canali del potassio può costituire un vantaggio: essa fa sì, infatti, che siano responsabili della maggior parte del «rumore» elettrico, derivante dalle variazioni casuali nel flusso della corrente ionica attraverso la membrana. L'esame dello spettro dei rumori in differenti tipi di membrana eccitabile è un campo di studio praticamente nuovo e promette di fornire importanti informazioni su alcune caratteristiche dei canali per gli ioni, altrimenti non osservabili. Affidati a Charles F. Stevens della Yale University School of Medicine, Franco Conti del Consiglio nazionale delle ricerche italiano e ad altri ricercatori, gli studi sui rumori hanno già fornito le migliori stime di cui si dispone per la conduttanza di un singolo canale e per il numero di canali del potassio che sono presenti nell'assone gigante di calamaro. Essi hanno dimostrato che esistono solo 50 canali del potassio circa per micrometro quadrato di membrana dell'assone o, grosso modo, un canale del potassio ogni 10 canali del sodio. La relativa scarsezza dei canali del potassio riduce ulteriormente la probabile entità della corrente che stabilisce una barriera per tale elemento.

Questo è tutto per quanto riguarda i risultati delle misurazioni elettrofisiologiche. Essendo oggi chiaro che i cambiamenti nelle conduttanze ioniche della membrana nervosa, dipendenti dalle differenze di potenziale sono determinati da ridistribuzioni strutturali all'interno dei canali ionici, sorge il problema se essi potrebbero essere individuabili con tecniche diverse da quelle elettriche. Tenendo presente questo punto, numerosi ricercatori, me compreso, hanno ricercato cambiamenti transitori in parecchi parametri che potrebbero essere associati con la configurazione molecolare delle particelle che costituiscono la barriera. Questi parametri includevano la birifrangenza della membrana, la diffusione della luce da parte dell'assone, l'emissione e l'assorbimento di luce dopo trattamento della membrana con un colorante fluorescente e la produzione e l'assorbimento di calore da

In un assone gigante di calamaro, l'autore insieme a George Malachowski e Dirk van Heiden della Marine Biological Association di Plymouth ha registrato una «famiglia» di correnti di sbarramento. Il potenziale di membrana è stato tenuto a − 100 millivolt e quindi portato mediante impulsi a − 120, − 80, − 40, 0 e + 40 millivolt. Quando ha raggiunto valori positivi, la corrente di sbarramento ha cominciato a fluire verso l'esterno.

parte dell'assone. Benché, in ogni caso, gli effetti si siano potuti osservare durante il passaggio degli impulsi nervosi o di impulsi dovuti alla tecnica di *voltage clamp*, si è dovuto concludere con rincrescimento che essi derivavano sempre da una risposta della membrana come un tutto unico e non erano correlati in maniera specifica a eventi che si verificavano all'interno dei canali ionici stessi.

Per giungere a una comprensione completa del funzionamento dei canali ionici, occorre studiarne la biochimica. Si può ragionevolmente ammettere che essi - alla stessa stregua di altre grosse molecole interessate nel trasporto di ioni attraverso membrane cellulari - siano grosse proteine, le quali avrebbero una sezione centrale idrofoba, piazzata nell'ambiente oleoso costituito dalle code degli idrocarburi che fanno parte dei fosfolipidi della membrana, e porzioni idrofile che sporgono nel mezzo acquoso su ogni lato della membrana. Come con altre proteine delle membrane, questa duplice natura rende difficoltosi l'isolamento e la purificazione della proteina per metodi di analisi strutturale com'è, per esempio, la cristallografia con i raggi X: infatti il processo di estrazione dalla membrana altera probabilmente la conformazione della proteina, inattivandola. Per poter usare tecniche di purificazione quali la cromatografia per affinità è inoltre necessaria una molecola adatta alla quale la proteina del canale si legherà in maniera selettiva. La tetrodotossina sembrerebbe la candidata favorita, ma fino a poco tempo fa, quando Michel Lazdunski dell'Università di Nizza è riuscito a prepararne dei derivati fluorescenti, non sembrava che la molecola fosse facile da manipolare chimicamente.

La chiave per il successo in ricerche di questo tipo consiste spesso nel trovare un materiale di partenza già molto ricco della proteina desiderata, come nel caso dell'organo elettrico della torpedine, utilizzato da Jean-Pierre Changeux dell'Istituto Pasteur di Parigi quale fonte di una proteina recettrice dell'acetilcolina. Nel frattempo, un altro tipo di organo elettrico è stato individuato in *Electrophorus*. Benché le elettroplacche di quest'ultimo non siano così ricche di canali del sodio come quelle della torpedine lo sono di recettori per l'acetilcolina, non sono poi così negative sotto questo aspetto; inoltre un grosso esemplare, fornisce un chilogrammo o più di tessuto elettrico. Partendo da studi sulla perdita della capacità di legare la tetrodotossina, riscontrata in preparati ottenuti da microsomi di canali del sodio, bombardati con elettroni a elevata energia, i miei colleghi Clive Ellory e Levinson sono giunti a un valore di 230 000 dalton per il peso molecolare della proteina che costituirebbe i canali. Lavorando con Michael A. Raftery al California Institute of Technology, Levinson e collaboratori sono riusciti a estrarre dall'organo elettrico di *Electrophorus* una proteina che lega la tetrodotossina e che è praticamente pura. Il suo peso molecolare è dello stesso ordine di grandezza. Pur essendovi ancora molti non indifferenti ostacoli tecnici da superare, gli elettrofisiologi possono sperare di riuscire a studiare, un giorno o l'altro, canali del sodio incorporati in membrane artificiali. Cominciano dunque a emergere vaghi lineamenti dell'organizzazione dei canali ionici in membrane eccitabili. L'ulteriore chiarimento dell'esatta struttura dei canali e della loro modalità di funzionamento risolverà quello che è, indubbiamente, uno dei più importanti problemi ancora insoluti riguardo al meccanismo della eccitabilità biologica.

Questo modello multifasico dei canali del sodio è stato proposto dall'autore sulla base di recenti studi sulle correnti di sbarramento per questo ione. Esso mostra l'ipotetica sequenza di stati configurazionali attraverso i quali le subunità che funzionano da barriere passano durante l'apertura e la chiusura dei canali. In risposta a uno spostamento verso valori positivi del potenziale della membrana dell'assone, le subunità che fungono da barriera subiscono una modificazione nella loro configurazione che li porta a uno stato attivato. Esse sono allora in grado di interagire in successione e, quando tutte e tre le subunità che controllano un dato canale lo hanno fatto, il canale si apre e fa passare gli ioni sodio. Dopo essere rimasto aperto per un breve periodo, il canale si richiude grazie a un altro cambiamento configurazionale che lo riporta allo stato inattivato. Le doppie frecce tra i vari stati indicano che ogni reazione può procedere nei due sensi. In pratica, però, la maggior parte dei canali ionici passa, attraverso l'intero ciclo, dallo stato di riposo a quello inattivato e di nuovo allo stato di riposo e questo avviene a ogni impulso. Una corrente di sbarramento misurabile accompagna soltanto la fase che passa dallo stato di riposo allo stato attivato o viceversa.

Le metastasi tumorali

Nella maggior parte dei casi, non è il tumore primitivo a causare la morte ma le sue metastasi dovute a cellule, particolarmente maligne, che sono in grado di spostarsi nell'organismo e originare nuovi tumori

di Garth L. Nicolson

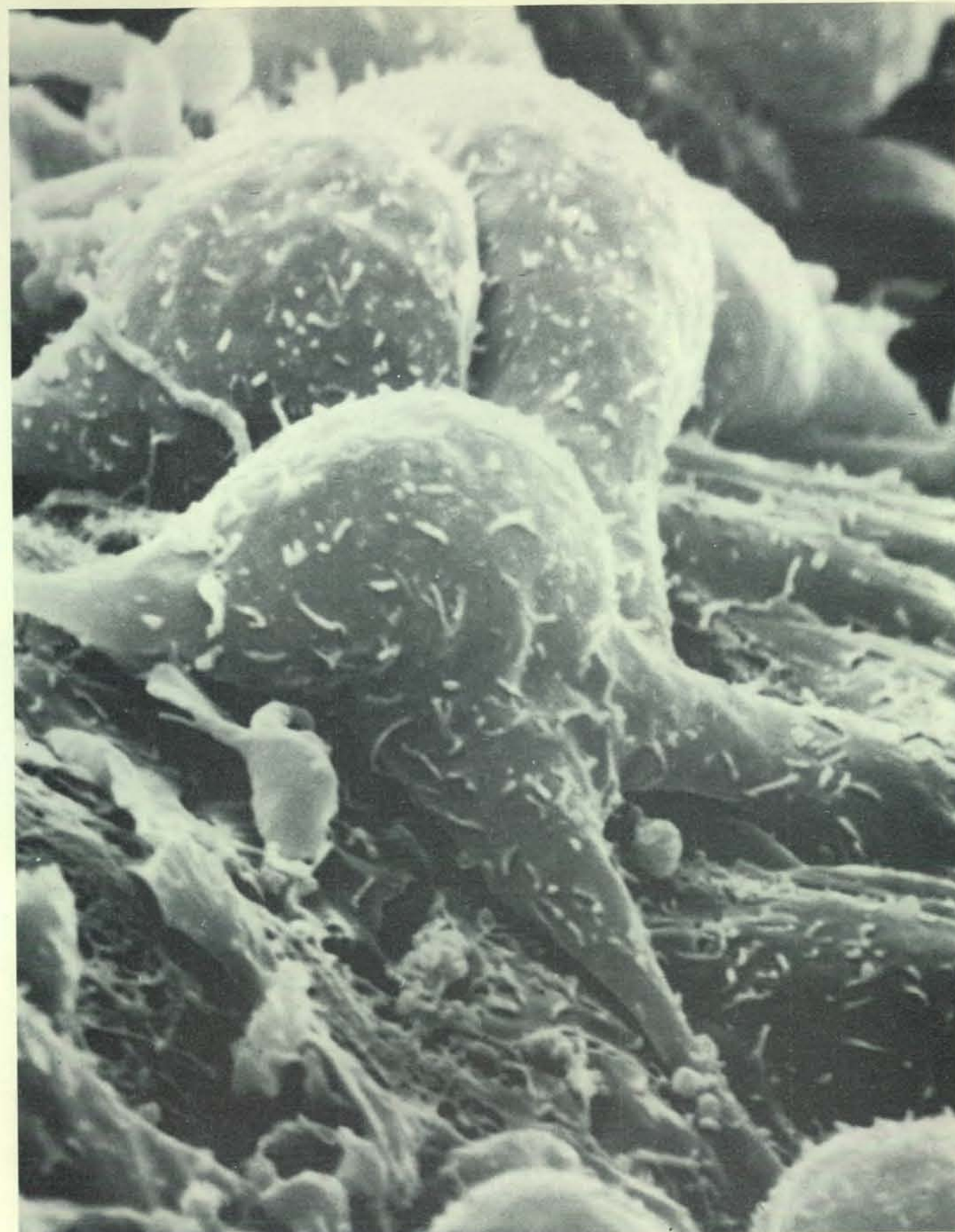
La maggior parte dei pazienti affetti da tumore maligno non muore per il tumore primitivo, ma per le sue metastasi: colonie tumorali multiple e diffuse derivate da cellule maligne che si staccano dal tumore originario e si spostano nell'organismo, spesso in sedi lontane. Se il tumore primitivo viene diagnosticato abbastanza precocemente può essere eliminato mediante la chirurgia, la radioterapia, la chemioterapia o una combinazione di questi trattamenti. Sfortunatamente, le colonie metastatiche sono più difficili da scoprire e da estirpare ed è spesso impossibile trattarle tutte con successo. Da un punto di vista clinico, quindi, la metastasi può essere considerata come l'evento terminale della storia naturale del tumore maligno la cui caratteristica è proprio la capacità di produrre metastasi.

Un tumore si origina dalla trasformazione di una cellula normale in una cellula che sfugge ai normali controlli dell'ospite sulla crescita e sulla differenziazione. La cellula trasformata prolifera dando luogo a un tumore, che può essere benigno o maligno. Un tumore è di solito considerato benigno se mantiene una struttura analoga a quella del tessuto da cui deriva, se cresce lentamente per semplice espansione, se rimane incapsulato da uno strato di tessuto connettivo e se i nuclei delle sue cellule si dividono quasi normalmente, con poche anomalie cromosomiche. Un tumore maligno, invece, possiede di solito una struttura tissutale atipica, cresce rapidamente e non rimane incapsulato, mostra numerose divisioni del nucleo e cromosomi anomali e invade il tessuto normale circostante, disseminando cellule che possiedono la capacità di colonizzare nuove sedi. Dal momento che le distinzioni tra condizione benigna e maligna basate sul tasso di crescita e sul grado di «anormalità» non sono assolute, il tumore maligno, o cancro, viene diagnosticato in base alla capacità di invadere i tessuti circostanti e di colonizzare sedi anche lontane. La comprensione della biologia delle metastasi tumorali è dunque di primaria importanza sia per capire la natura del tumore maligno, sia per prolungare o salvare la vita dei pazienti affetti da tumore maligno distruggendo quelle cellule tumorali metastatiche che sono le più pericolose per la sopravvivenza.

Il processo della metastatizzazione tumorale comprende una serie complessa di eventi che implicano interazioni tra le stesse cellule maligne derivate dal tumore primitivo e tra queste cellule e l'organismo. Irving Zeidman della University of Pennsylvania School of Medicine ha stabilito i passaggi sequenzali che portano alla colonizzazione di un tumore metastatico distante. Essi sono: 1) estensione ai tessuti circostanti; 2) penetrazione nelle cavità e nei vasi del corpo; 3) rilascio di cellule tumorali per la migrazione in altre sedi; 4) reinvasione del tessuto nel punto di arresto e 5) trasformazione del nuovo ambiente per favorire la sopravvivenza della cellula tumorale e quindi la vascolarizzazione e la crescita del tumore.

Dapprima le cellule maligne si espandono dal luogo di origine, invadendo i tessuti normali circostanti. Alcune di queste cellule penetrano in una cavità del corpo (come la cavità toracica o quella addominale) o nel sangue, nella linfa, o nel liquido cerebro-spinale e vengono liberate. Nel sangue o nei vasi linfatici singole cellule o gruppi di cellule (chiamati emboli tumorali) possono essere trasportati in sedi molto lontane dal tumore primitivo. Le cellule che sopravvivono al trasporto nel sistema linfatico vengono di solito arrestate nei linfonodi regionali: quelle che sopravvivono nel flusso ematico finiscono con l'arrestarsi nelle venule o nei capillari. Le cellule maligne, tuttavia, possono sfuggire all'ambiente ostile della linfa o del sangue infiltrando i tessuti adiacenti. Infine le micrometastasi si vascolarizzano e la rete di vasi sanguigni di nuova formazione fornisce le sostanze nutritive che permettono alle cellule di proliferare e di formare una massa tumorale più grande. Durante l'intero processo metastatico le cellule maligne devono continuamente difendersi dall'attività distruttiva dei diversi meccanismi antitumorali dell'ospite.

Sono state proposte diverse teorie per spiegare il primo evento della metastasi, l'invasione dei tessuti adiacenti. George Eaves della University of Leeds School of Medicine ha proposto che il comportamento invasivo possa essere spiegato in gran parte sulla base di considerazioni di ordine meccanico. Iniettando resina dentaria a presa rapida in campioni di tessuto, si otteneva l'infiltrazione delle zone a minore resistenza, un comportamento che ricorda quello di molti tumori maligni. Tuttavia, non bastano semplici modelli meccanici per spiegare il comportamento invasivo estremamente rapido di alcuni tumori; esperimenti condotti su colture di tessuto mostrano chiaramente che le cellule maligne altamente invasive riescono a penetrare in uno strato di cellule normali anche in assenza di un qualsiasi stimolo meccanico. Dale R. Coman della University of Pennsylvania School of Medicine ha suggerito che sia la motilità delle cellule maligne a renderle capaci di infiltrare i tessuti circostanti, ma ci sono diverse cellule normali che sono notevolmente mobili, ma non invasive.

La presenza nel tessuto tumorale di alte concentrazioni di determinati enzimi ha suggerito che il comportamento invasivo di cellule tumorali possa essere facilitato dalla distruzione enzimatica della normale matrice connettivale esistente tra le cellule. A. R. Poole e i suoi collaboratori dello Strangeways Research Laboratory hanno messo a confronto gli enzimi proteolitici secreti da colture di cellule provenienti da tumori umani benigni della mammella; Bengt Sylvén ha condotto esperimenti analoghi presso l'Istituto Karolinska in Svezia. Poole ha scoperto che una proteinasi, la catepsina B, mostrava un tasso di accumulo più elevato nelle colture di cellule maligne. Tuttavia altri studi, hanno dimostrato poca differenza nella quantità di tali enzimi rilasciati da tessuto tumorale e da tessuto normale, per cui

Questa microfotografia eseguita dall'autore mostra cellule metastatiche di melanoma, ingrandite 7000 volte, mentre infiltrano la parete di un vaso sanguigno. Le cellule di melanoma di topo che metastatizzano tipicamente per via ematica sono state incubate con un lembo di tessuto venoso e quindi mantenute a contatto con quest'ultimo in una coltura di laboratorio. Le cellule tumorali hanno aderito alle cellule endoteliali che rivestono la superficie interna della vena e hanno spinto i propri prolungamenti in fessure che si aprono al ritrarsi delle cellule endoteliali sottostanti. L'adesione endoteliale e l'infiltrazione sono due stadi del processo che consente alla cellula tumorale circolante di formare una nuova colonia.

l'invasività non può essere spiegata soltanto sulla base dell'azione enzimatica. Essa dipende probabilmente da una combinazione di vari processi: meccanici, enzimatici e di motilità cellulare.

Estendendosi dal tumore originario, le cellule maligne devono staccarsi da esso per poter metastatizzare. Leonard Weiss e George H. Poste del Roswell Park Memorial Institute ritengono che alcune cellule tumorali o cellule normali possano lisarsi, o scoppiare, aumentando il livello extracellulare di enzimi degradativi che liberano ammassi di cellule maligne. Coman ha trovato, per esempio, che agitando aggregati di cellule maligne provenienti da un tumore della mammella si separavano più cellule rispetto a quelle che si separavano se gli aggregati erano stati ricavati da tessuto normale della mammella.

La liberazione di cellule maligne nel sangue o nella linfa può portare a metastasi tumorali diffuse. Zeidman, Bernard Fisher ed Edwin R. Fisher della University of Pittsburgh School of Medicine hanno dimostrato che, siccome i sistemi sanguigno e linfatico sono collegati, le cellule tumorali possono passare dai vasi linfatici ai vasi sanguigni e viceversa. I Fisher hanno marcato cellule di un carcinoma (tumore maligno di natura epiteliale) con cromo radioattivo, iniettandole nelle vene periferiche di ratti. Entro un'ora dall'iniezione, nei vasi linfatici si potevano rinvenire cellule tumorali vive e l'iniezione di linfa in altri animali produceva carcinomi. Questi risultati dimostrano che le cellule maligne in circolo non vengono sempre intrappolate nel primo capillare che incontrano, ma sono in grado di superare questa barriera e di passare dal sangue nel sistema linfatico, probabilmente attraverso i linfonodi.

Durante il suo sviluppo, un tumore maligno libera probabilmente in modo continuo nella linfa o nel sangue cellule maligne. Pietro M. Gullino del National Cancer Institute ha controllato tumori mammari trapiantati, capaci di abbandonare i vasi sanguigni, in animali da esperimento e ha trovato che, quando un tumore è cresciuto di alcuni grammi, è in grado di mettere in circolazione milioni di cellule maligne al giorno. La presenza di cellule maligne nel sangue (o nella linfa) tuttavia non sempre conduce alla formazione di metastasi lontane. A. J. Salsbury del Brompton Hospital di Londra è stato uno dei primi a sottolineare il fatto che la presenza nel sangue di cellule tumorali liberate da un tumore solido primitivo non significa necessariamente che un paziente debba in seguito soccombere per metastasi multiple. Il fatto è che la maggioranza delle cellule maligne liberate nel sangue muore molto rapidamente. Solo alcune sembrano in grado di sopravvivere nell'agitato e ostile ambiente del torrente circolatorio.

Isaiah J. Fidler e Zeidman hanno analizzato il destino delle cellule di melanoma nei topi. (Il melanoma è un tumore maligno che si sviluppa sulla pelle degli animali e dell'uomo; spesso è fortemente pigmentato per la presenza di melanina). Fidler e Zeidman iniettarono in una vena periferica cellule di melanoma di topo il cui DNA era stato marcato con iodio radioattivo. Essi scoprirono che, entro alcuni minuti, la maggior parte delle cellule iniettate veniva arrestata nel polmone, sede del primo sistema capillare che le cellule incontravano. La maggior parte delle cellule tumorali moriva in quella sede; quelle che sopravvivevano, invece, continuavano a circolare. Entro un giorno, tuttavia, solo l'uno per cento delle cellule iniettate era ancora vivo. Dopo due settimane, quando le metastasi del melanoma potevano già essere osservate nel polmone, solo un decimo dell'uno per cento delle cellule marcate originariamente iniettate era ancora vivo. I risultati indicavano che le metastasi di origine ematica devono costituire un processo altamente selettivo e che solo una percentuale molto piccola delle cellule maligne in grado di penetrare nel torrente ematico e circolare in esso sopravvive per formare metastasi.

Mentre le cellule maligne circolano nel sangue esse contraggono una quantità di importanti rapporti cellulari, come è stato dimostrato nel mio laboratorio a Irvine, presso l'Università di California e nei laboratori di Fidler (che si trova ora al National Cancer Institute's Frederick Cancer Research Center), Gabriel J. Gasic e Tatiana Gasic della University of Pennsylvania School of Medicine e Lance A. Liotta del National Cancer Institute. Queste interazioni sono verosimibilmente in rapporto con la sopravvivenza delle cellule con la loro deposizione nei capillari. Le cellule maligne circolanti possono, per esempio, aderire ad altre cellule tumorali formando aggregati. Ma esse possono aderire anche a cellule normali come linfociti e piastrine. E quando entrano in un letto capillare, possono aderire alle cellule endoteliali che rivestono i vasi. Più grande è l'aggregato, maggiore è la probabilità che esso si arresti nei capillari e formi nuove colonie tumorali. Una volta fermate nei capillari, alcune cellule tumorali di origine ematica stimolano la produzione di fibrina, la proteina fibrosa che è normalmente coinvolta nella formazione di coaguli sanguigni. B. A. Warren della University of Western Ontario ha trovato che la rete di fibrina risultante tende a proteggere le cellule maligne e le mette in grado di proliferare.

Quando le cellule maligne si sono arrestate nelle venule o nei capillari, devono attraversare la parete dei vasi e penetrare nei tessuti extravasali per poter sopravvivere. Questa fase dell'infiltrazione è stata studiata da Randall Kramer nel mio laboratorio. Si è osservato che le cellule maligne inducono una modificazione radicale nella capacità di coesione delle cellule endoteliali (che rivestono la superficie interna dei vasi sanguigni) alla loro matrice extracellulare sottostante: le cellule endoteliali si ritraggono, lasciando spazi liberi attraverso i quali le cellule tumorali possono uscire dal sistema vascolare. Una volta che le cellule tumorali hanno invaso il tessuto extravascolare, devono formarsi un nuovo microambiente. La crescita della nuova colonia tumorale è limitata fino a che non si rende disponibile una nuova fonte di rifornimento ematico per il trasporto di elementi nutritivi e il drenaggio dei prodotti di rifiuto delle cellule. La vascolarizzazione del nuovo tumore è stimolata da «fattori angiogenetici tumorali» rilasciati dalle cellule maligne (si veda l'articolo *La vascolarizzazione dei tumori* di Judah Folkman in «Le Scienze», n. 97, settembre 1976). Questi fattori sembrano funzionare stimolando in modo specifico la proliferazione delle cellule endoteliali attraverso la divisione cellulare e il movimento delle cellule endoteliali per formare nuovi capillari.

Il processo della metastatizzazione, dunque dipende da numerose interazioni tra cellule maligne e cellule normali dell'ospite. Probabilmente solo poche cellule maligne di un tumore primitivo sono in grado di avvantaggiarsi da tali interazioni e quindi di completare tutti i passaggi della sequenza metastatica. Solo una piccola frazione delle cellule provenienti dal tumore primitivo invadono una cavità del corpo o entrano nel sistema sanguigno o linfatico e si distaccano: di queste cellule solo alcune sopravvivono allo spostamento e un numero ancora più piccolo sopravvive all'intero processo di metastatizzazione a ogni passaggio del quale sono messi in atto svariati meccanismi di difesa dell'organismo.

Il fatto che solo una piccola frazione delle cellule tumorali sopravviva per formare nuove colonie metastatiche suggerisce che le cellule che riescono a sopravvivere devono presentare caratteristiche particolari. Forse esse costituiscono una sottopopolazione di cellule tumorali in grado di permettere la metastatizzazione? Le difese dell'ospite possono essere responsabili della selezione naturale di una simile sottopopolazione di cellule particolarmente maligne, con caratteristiche altamente metastatiche?

Questa ipotesi può essere verificata in tumori maligni metastatici, tra cui uno dei più utili è il melanoma murino. Il melanoma metastatico è uno dei tumori maligni dell'uomo meno suscettibili al trattamento - in gran parte a causa della frequenza delle metastasi di origine linfatica ed ematica - e i melanomi non umani hanno le stesse caratteristiche. Il melanoma murino *B*16, fatto sviluppare in un ceppo di topi neri allevati presso il Jackson Laboratory, può essere trapiantato da un topo all'altro e può essere fatto crescere anche in coltura. Quando le cellule del melanoma *B*16 vengono innestate nella pelle del topo, il melanoma di solito metastatizza a un tasso da basso a medio. Allo scopo di ottenere un dosaggio più preciso

CAPSULA CONNETTIVALE

Il tumore è una massa di tessuto che si origina dalla proliferazione di una cellula trasformata in modo da sfuggire ai normali controlli sulla crescita e la differenziazione. In generale un tumore benigno *(a sinistra)* appare quasi normale per quel che riguarda la struttura istologica, si accresce lentamente e rimane incapsulato all'interno del tessuto connettivo; esso non infiltra i tessuti circostanti e le sue cellule non si staccano e non metastatizzano. Un tumore maligno *(a destra)*, invece, è atipico nella struttura, di solito si accresce più rapidamente e non rimane incapsulato entro il tessuto connettivo. Esso s'infiltra nel tessuto circostante e generalmente alcune cellule si staccano e metastatizzano.

La diffusione metastatica di un melanoma maligno, un tumore pigmentato della pelle, può iniziare quando cellule maligne invadono un vaso linfatico e vengono trasportate nella linfa sino a un linfonodo. Le cellule, momentaneamente arrestate, possono penetrare in un vaso sanguigno e quindi essere trasportate nel torrente circolatorio, attraverso il cuore, verso altre regioni del corpo. Le metastasi più frequenti si formano nel cervello, nei polmoni, nel fegato e nelle ovaie.

del potenziale metastatico, quantità note di cellule tumorali vengono di solito iniettate direttamente nel sistema circolatorio attraverso la vena caudale. Le cellule circolano, si depositano, invadono, sopravvivono e si moltiplicano formando alcune caratteristiche colonie tumorali pigmentate nei polmoni e in numero inferiore nelle ovaie, nelle cavità del corpo, nel fegato e nel cervello.

Fidler, lavorando alla School of Dental Medicine della Università della Pennsylvania, eseguì un esperimento nel quale manipolò cellule di melanoma *B*16 allo scopo di selezionare sottopopolazioni dotate di una aumentata capacità di colonizzare i polmoni. Egli cominciò rimuovendo chirurgicamente le colonie del melanoma che si formavano nei polmoni dei topi in cui le cellule *B*16 erano state iniettate per via endovenosa. Egli rese le cellule adatte a crescere in un ambiente colturale e quindi utilizzò le cellule coltivate, indicate come ceppo cellulare variante di melanoma *B*16-*F*1, per iniettarle per via endovenosa in un altro gruppo di topi. Le nuove colonie sperimentali cancerogeniche verso il polmone che risultarono da questa seconda selezione, vennero messe, a loro volta, in coltura per ottenere un nuovo ceppo cellulare, *B*16-*F*2. La sequenza: iniezione, colonizzazione polmonare, raccolta della coltura tissutale per la reiniezione nei topi veniva ripetuta continuamente e, a ogni ciclo successivo, l'efficienza con cui i ceppi varianti venivano arrestati nei polmoni, sopravvivevano, infiltravano i tessuti e davano origine a colonie tumorali nel polmone aumentava regolarmente.

Dopo 10 cicli Fidler definì un ceppo variante di melanoma, *B*16-*F*10, che formava un numero significativamente maggiore di colonie tumorali nel polmone rispetto al ceppo parentale *B*16 o al ceppo selezionato *B*16-*F*1, se un dato numero di cellule maligne venivano iniettate nei topi. Questa famiglia di ceppi varianti di cellule tumorali che posseggono differenti attività metastatiche si è dimostrata assai utile in una quantità di studi biochimici e biologici sul processo di metastatizzazione tumorale. Il suo successo come modello di metastasi è probabilmente dovuto al fatto che l'originale linea tumorale *B*16 non consisteva di una popolazione cellulare uniforme: era probabilmente eterogenea in rapporto alle caratteristiche della particolare cellula tumorale necessaria per la metastasi.

Sono qui illustrati i vari stadi di una metastasi di origine ematica. Il tumore primitivo si espande per la la proliferazione delle cellule maligne, ma si accresce rapidamente solo dopo che la vascolarizzazione fornisce un adeguato rifornimento ematico *(1)*. Il tumore invade il tessuto circostante *(2)*. Alcune cellule, singolarmente o in gruppo, attraversano la membrana basale di un capillare, si comprimono tra le cellule endoteliali ed entrano nel torrente circolatorio, dove possono interagire con cellule del sangue come i linfociti *(3)*. Le cellule tumorali sono trasportate attraverso il cuore fino ai capillari polmonari entro cui si arrestano; qui le cellule tumorali infiltrano la parete del capillare *(4)*, entrano nel tessuto alveolare dove proliferano dando luogo alla formazione di una nuova colonia tumorale *(5)*.

Le osservazioni cliniche indicano che esiste una tendenza, in particolari tipi di tumori, a metastatizzare preferenzialmente entro organi determinati. Per esempio, i carcinomi della mammella tendono a diffondersi al cervello o ai polmoni; quelli polmonari spesso formano metastasi al cervello e alle ghiandole surrenali, mentre il carcinoma della prostata forma in genere metastasi nelle ossa. Comportamenti selettivi nella metastasi sono stati osservati anche in tumori non umani. Se lo schema della distribuzione metastatica per un dato tipo di cellula non è casuale, quale fenomeno può spiegarla?

Un'ipotesi fu avanzata all'inizio del secolo dai ricercatori James Ewing e Stephen Paget. Si riteneva che la distribuzione metastatica fosse determinata sia da condizioni anatomiche e vascolari, sia dal microambiente che si veniva a creare nella sede della metastasi e che consentiva la sopravvivenza del tumore e la sua crescita. L'implicazione era che le cellule tumorali crescono rigogliosamente in un «terreno» particolare se sono in grado di adattarvisi, per cui tale ipotesi venne indicata come «ipotesi del terreno». Più recentemente si è chiarito che le proprietà delle cellule maligne stesse influenzano grandemente il loro schema di metastasi. Come hanno messo in rilievo Everett. D. Sugarbaker e Alfred S. Ketchem della University of Miami School of Medicine, la distribuzione e la sopravvivenza delle singole cellule maligne dipende probabilmente da numerosi fattori, che comprendono le proprietà specifiche delle cellule stesse, l'ambiente dove esse vivono e crescono negli organi specifici e la condizione di immunità antitumorale presentata dall'ospite.

In collaborazione con Fidler abbiamo studiato la specificità con cui la sottopopolazione delle cellule di melanoma di topo *B*16 si insediano nel polmone e lo colonizzano. Le cellule tumorali vengono intrappolate nel polmone semplicemente perché è il primo letto capillare che incontrano? O invece vanno in cerca del tessuto polmonare? Abbiamo iniettato le cellule *B*16-*F*1 relativamente non specifiche e le cellule *B*16-*F*10 sia nella vena caudale (dalla quale avrebbero incontrato per primi i capillari del polmone) e nel ventricolo sinistro del cuore - che nel sistema circolatorio si trova a valle rispetto al polmone - in modo che le cellule avrebbero incontrato altri letti capillari prima di raggiungere i polmoni. La distribuzione iniziale delle cellule tumorali marcate con elementi radioattivi era notevolmente diversa, suggerendo che esse erano state sicuramente arrestate nei capillari a valle della sede di iniezione. Tuttavia entro un giorno dall'iniezione, la distribuzione aveva caratteri simili e il numero di cellule sopravvissute di ciascun ceppo era lo stesso, indipendentemente dal punto di ingresso nel sistema circolatorio. Alla fine di due settimane fu trovato un numero pressoché uguale di tumori polmonari derivati da entrambe le vie, in cui le cellule *F*10, come ci si attendeva, avevano formato un numero di tumori significativamente maggiore delle cellule *F*1. I risultati indicavano che le cellule tumorali destinate a formare metastasi in un organo particolare possono anche staccarsi da una sede iniziale di arresto e farsi trasportare nel sangue fino a quando trovano l'ambiente adatto per depositarsi, sopravvivere e crescere.

La specificità d'organo nella colonizzazione non sembra dunque significare solamente che alcune cellule invadono preferibilmente un organo particolare, ma anche che esse diventano meno adatte a colonizzare altri organi. In questo esperimento e in molti altri abbiamo trovato che, mentre le cellule *B*16-*F*1 formavano tumori fuori dal polmone del topo (soprattutto nei linfonodi, nelle ovaie, nella cavità toracica, nel fegato, negli intestini, nel cervello e nelle ghiandole surrenali), le cellule *F*10 altamente selezionate non li formavano. Esse colonizzavano soltanto il polmone.

Nel mio laboratorio Kenneth W. Brunson ottenne un ulteriore progresso nel processo selettivo: egli cominciò con il ceppo cellulare *B*16-*F*1 che colonizza il polmone in modo predominante ma non esclusivo, e da esso selezionò sottopopolazioni che colonizzano preferibilmente zone del tutto diverse. Nel corso di un esperimento, egli asportò tumori cerebrali che sono raramente prodotti dal ceppo cellulare *B*16-*F*1, riuscì a coltivarne le cellule e ottenne un ceppo da lui chiamato *B*16-*B*1. Quindi egli utilizzò il procedimento che Fidler aveva sviluppato per ottenere il ceppo *F*10 specifico per il polmone, ma a ogni passaggio asportava i tumori cerebrali. Dopo 10 selezioni

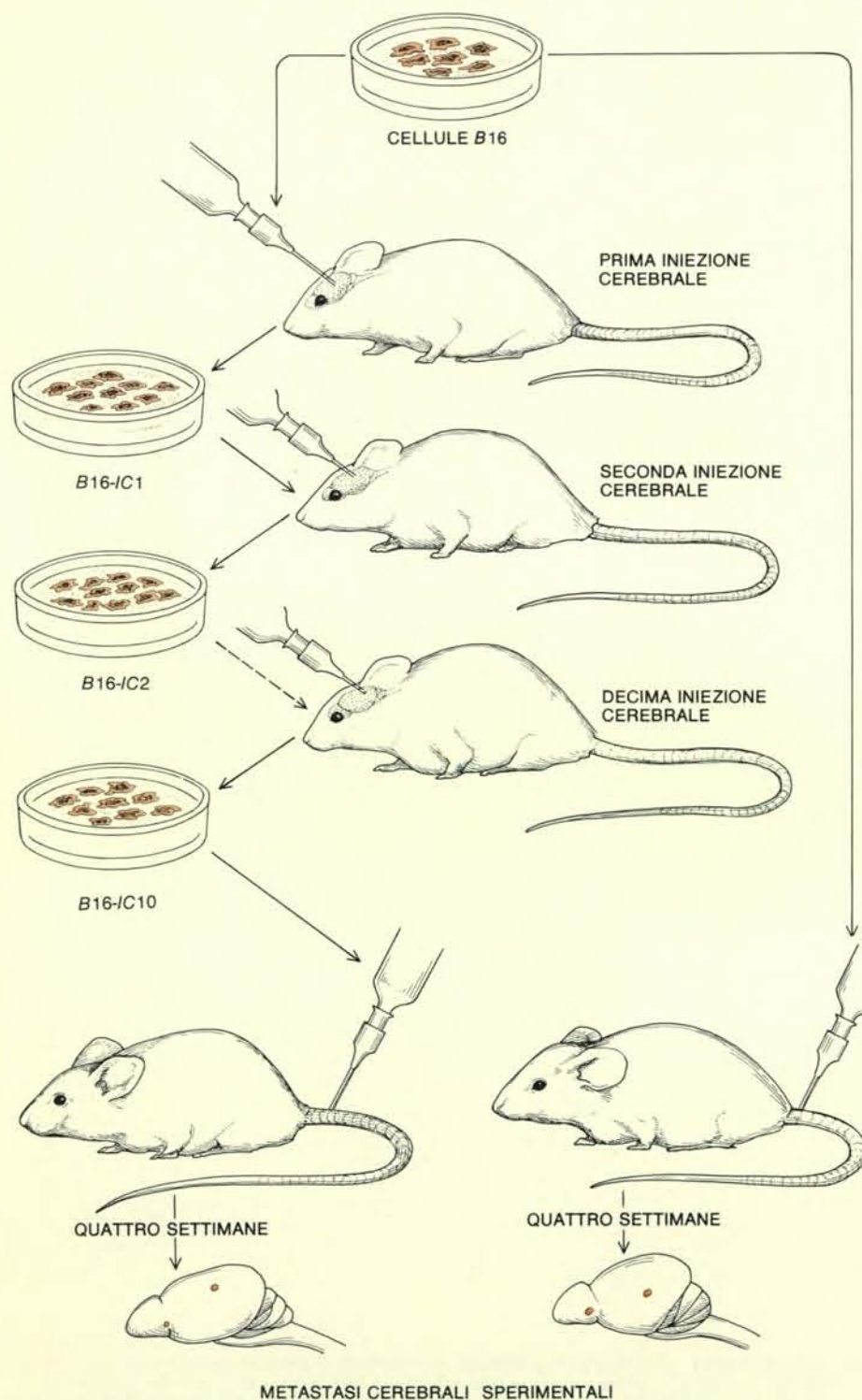
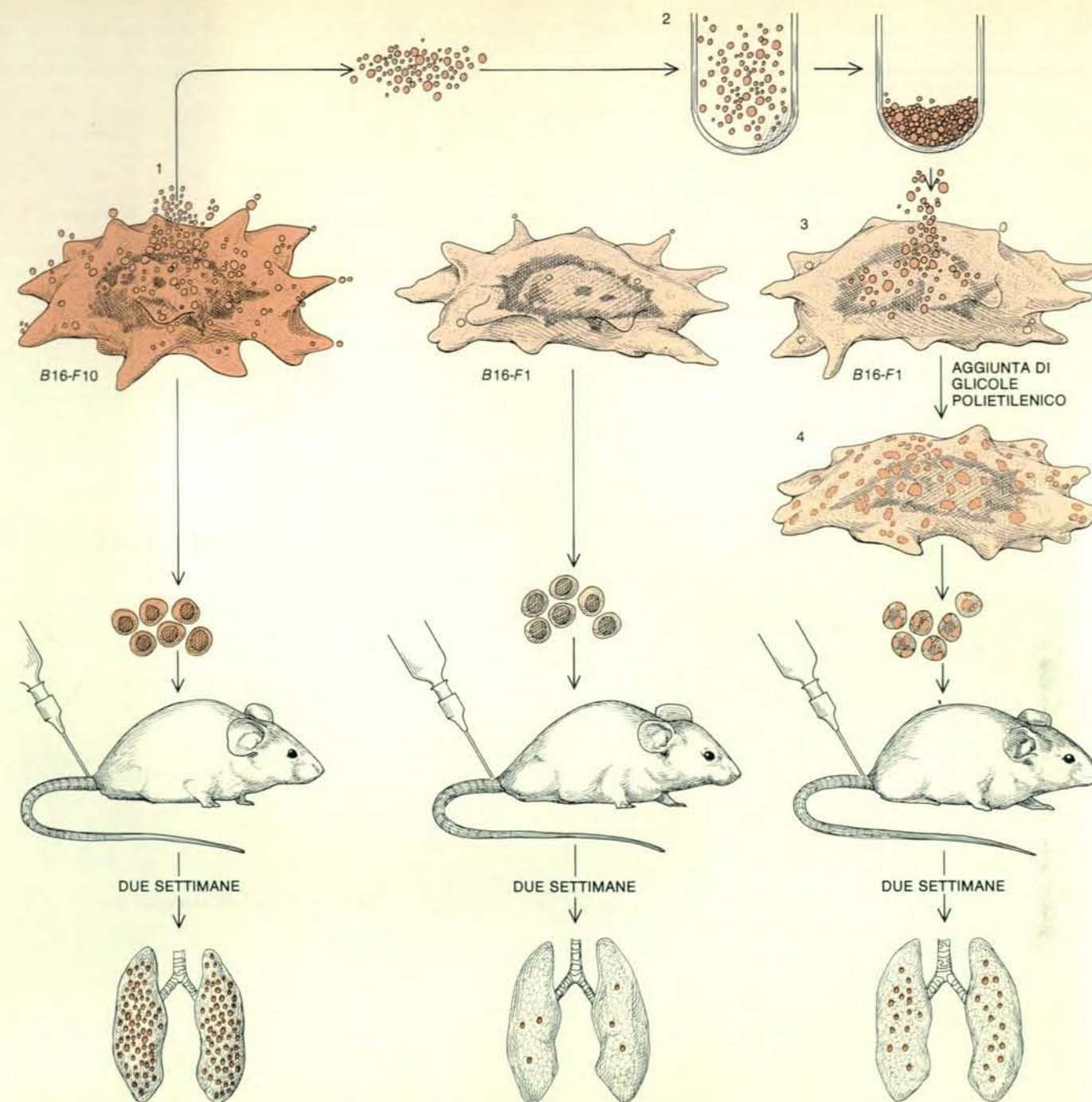
La selezione di ceppi di cellule di melanoma ad aumentata capacità di metastatizzazione nei polmoni *(a sinistra)* fu realizzata da Isaiah J. Fidler. Kenneth W. Brunson, del laboratorio dell'autore, selezionò ceppi che formano metastasi in una ragione particolare del cervello *(a destra)*. Nel caso della selezione per il polmone, Fidler prese cellule in coltura di melanoma *B*16 ricavate dal tumore di un topo, le iniettò nella vena caudale di topi e raccolse le cellule dalle risultanti colonie tumorali polmonari. Queste cellule *B*16-*F*1 vennero a loro volta messe in coltura e nuovamente iniettate per via endovenosa. Le cellule *B*16-*F*10 altamente selezionate prodotte nel corso di 10 di questi cicli formano metastasi polmonari sperimentali in numero significativamente superiore che non i ceppi *B*16 non selezionati e colonizzano soltanto i polmoni. Brunson iniettò cellule *B*16 nel ventricolo sinistro del cuore e prelevò i tumori cerebrali che ne derivarono. Le cellule tumorali, indicate come *B*16 - *B*1, furono nuovamente messe in coltura e iniettate nel cuore; questa volta i tumori vennero prelevati solo dalla scissura entorinale vicina alla parte anteriore del cervello *(nel rettangolo)*. Il procedimento fu ripetuto con le cellule iniettate per via endovenosa. Dopo 10 selezioni Brunson ottenne il ceppo *B*16-*B*10*n*, altamente metastatiche solo nella sede della scissura entorinale cerebrale.

Brunson aveva sviluppato un ceppo *B16-B10* che metastatizza in una zona particolare del cervello del topo, la fessura entorinale, tra la corteccia cerebrale e il bulbo olfattorio. Mediante un processo analogo, venne selezionata un'altra linea cellulare tumorale (*B16-O10*) che colonizza in modo preferenziale le ovaie, una zona che, nella specie umana, è frequentemente sede di metastasi causate da melanoma.

La colonizzazione preferenziale da parte dei ceppi cellulari *B16* di organi come il cervello, è il risultato dell'adattamento casuale delle cellule del melanoma all'ambiente particolare (come pretenderebbe l'ipotesi del terreno) o di proprietà specifiche possedute originariamente da determinate cellule tumorali? Brunson condusse esperimenti paralleli nei quali le cellule *B16* e successivamente i ceppi cellulari tumorali selezionati erano impiantati direttamente nel cervello e non iniettati in circolo eliminando in tal modo la fase di messa in circolo e di arresto nei capillari. Dopo 10 selezioni necessarie all'adattamento e alla crescita nel cervello, le cellule *B16* del melanoma recuperate dai tumori cerebrali non risultavano più efficaci in termini di metastasi cerebrali (se iniettate nella cute o nel sangue) rispetto al ceppo iniziale *F1*. In altre parole nella sequenza da *B1* a *B10n* non viene selezionata solo la pura adattabilità alla crescita nel cervello, ma anche la capacità di stabilirvisi e di colonizzare quella sede.

Il successo nell'identificazione delle proprietà specifiche dei ceppi cellulari tumorali che li rendono capaci di formare metastasi in sedi particolari potrebbe essere un primo grande passo verso lo sviluppo di un metodo capace di bloccare la metastasi nei pazienti cancerosi. Sebbene siamo ancora lontani dal sapere con precisione quali proprietà cellulari contribuiscono al successo della metastasi, è evidente che la superficie della cellula tumorale è direttamente coinvolta in fasi del processo quali l'arresto della cellula in una sede particolare.

George Poste e io abbiamo condotto un esperimento per dimostrare il ruolo svolto dalla superficie cellulare dopo aver trovato che le cellule *B16-F10* in coltura perdevano frammenti di membrana cellulare sotto forma di piccole vescicole chiuse. Raccogliemmo le vescicole e le purificammo mediante centrifugazione in provette contenenti un gradiente di una soluzione zuccherina densa. Le vescicole purificate vennero aggiunte a una coltura di cellule *B16-F1* insieme al polimero glicole polietilenico. In presenza del polimero le vescicole *F10* si fondevano e diventavano parte integrante della membrana cellulare della cellula *F1*. Quando le cellule *F1* arricchite con le vescicole della membrana di *F10* erano iniettate nei topi, esse metastatizzavano in maniera più simile a quella del ceppo altamente metastatico *B16-F10* che non a quella delle cellule normali *B16-F1*, indicando che la membrana cellulare è certamente coinvolta nella metastasi.

In che modo, allora, le membrane cellulari del ceppo *B16* differiscono tra loro, dato che colonizzano sedi diverse? Abbiamo messo a confronto le componenti proteiche delle superfici dei diversi ceppi cellulari incorporando molecole radioattive nelle proteine, dissolvendo le membrane cellulari in soluzioni detergenti e sottoponendo le proteine solubili della membrana a elettroforesi, in modo da separarle in base alle loro dimensioni e alla carica elettrica. Ciascuno dei tre ceppi del melanoma - quelli che colonizzano il polmone, il cervello e le ovaie - mostra una disposizione differente delle proteine specifiche di superficie e, in generale, la quantità (o l'accessibilità) di quelle specifiche proteine di superficie aumenta con il potenziale metastatico della cellula.

La colonizzazione preferenziale del cervello non è il risultato di un adattamento all'ambiente cerebrale. Ciò è stato dimostrato mediate un esperimento in cui cellule *B16* non furono iniettate nel sangue ma direttamente nel cervello. I melanomi che si formarono vennero asportati e le loro cellule messe in coltura e di nuovo iniettate nel cervello. Dopo 10 di tali cicli le cellule raccolte e iniettate in topi davano luogo a un numero di metastasi non superiore a quello delle cellule originarie del ceppo *B16* quando queste venivano iniettate per via endovenosa. Sembra che le cellule altamente metastatiche per il cervello appartengano a una sottopopolazione del ceppo *B16*.

È stato dimostrato che la membrana cellulare svolge un ruolo nel determinare la capacità metastatica delle cellule *B16*. Frammenti di membrana che erano stati spontaneamente dispersi sotto forma di vescicole *(1)* dalla membrana esterna delle cellule *B16-F10*, altamente metastatiche *(in colore più intenso)* furono raccolte dal mezzo di coltura, concentrate *(2)* e aggiunte a una coltura di cellule *B16-F1* a minore capacità di metastatizzazione *(3)*. In presenza di glicole polietilenico le vescicole *F10* si fusero con le mebrane delle cellule *F1 (4)*. I tre ceppi cellulari vennero iniettati in topi. Le cellule *F1*, le quali contenevano frammenti di membrana *F10*, davano origine a tumori polmonari in quantità significativamente maggiore delle cellule *F1* non trattate, sebbene non in quantità uguale a quella delle stesse cellule *F10*.

La maggior parte dei ricercatori ritengono che i tumori si sviluppino da un'unica cellula trasformata. Sottopopolazioni che possiedono proprietà diverse possono quindi crescere soltanto se alcune cellule di un tumore si evolvono indipendentemente anche se si trovano nello stesso ambiente. Un periodo prolungato di crescita tumorale nell'organismo umano (o la ripetuta diffusione dei tumori mediante trapianti in serie negli animali da esperimento) spesso porta a modificazioni progressive della struttura del tumore, delle proprietà antigeniche, delle caratteristiche metaboliche, delle proprietà di crescita e del potenziale metastatico.

Diversi ricercatori hanno richiamato l'attenzione sul fenomeno della progressione tumorale, tra cui Peter C. Nowell della University of Pennsylvania School of Medicine, Leslie Foulds del Chester Beatty Research Institute di Londra, George e Eva Klein dell'Istituto Karolinska, Emanuel Farber della Università di Toronto e H. J.-P. Ryser della Boston University School of Medicine. Quando una popolazione di cellule tumorali si

sviluppa lentamente, il tumore nel suo complesso si modifica e le sottopopolazioni che possiedono proprietà divergenti si evidenziano mentre la loro frequenza nella popolazione aumenta sotto la spinta della pressione selettiva da parte delle difese dell'ospite, portando alla fine alla comparsa di cellule con malignità aumentata. Le sottopopolazioni che possiedono potenziali metastatici differenti esistono come tali nel tumore primitivo, oppure l'intera popolazione si adatta semplicemente a un dato ambiente?

Un modo per distinguere tra queste due possibilità è stato fornito dal classico esperimento del 1943 in cui Salvador Luria e Max Delbrück riuscirono per primi a distinguere tra selezione e adattamento quale origine dei ceppi mutanti dei batteri. Fidler e Margaret L. Kripke ripeterono quell'esperimento con il ceppo non selezionato di melanoma *B*16 e con un ceppo di fibrosarcoma (tumore del tessuto connettivo) proveniente da un tumore indotto in un topo mediante irraggiamento ultravioletto; noi eseguimmo l'esperimento a Irvine con un ceppo di sarcoma di topo. In ciascun caso sospensioni di cellule tumorali non selezionate vennero divise in due frazioni. Una frazione fu iniettata per via endovenosa in un ampio gruppo di topi; l'altra frazione fu suddivisa in colture separate dalle quali venne isolata e fatta crescere una singola cellula per produrre un gran numero di cloni differenti, o sottopopolazioni di cellule, con una costituzione genetica identica. Questi cloni vennero inoltre iniettati per via endovenosa in gruppi di topi.

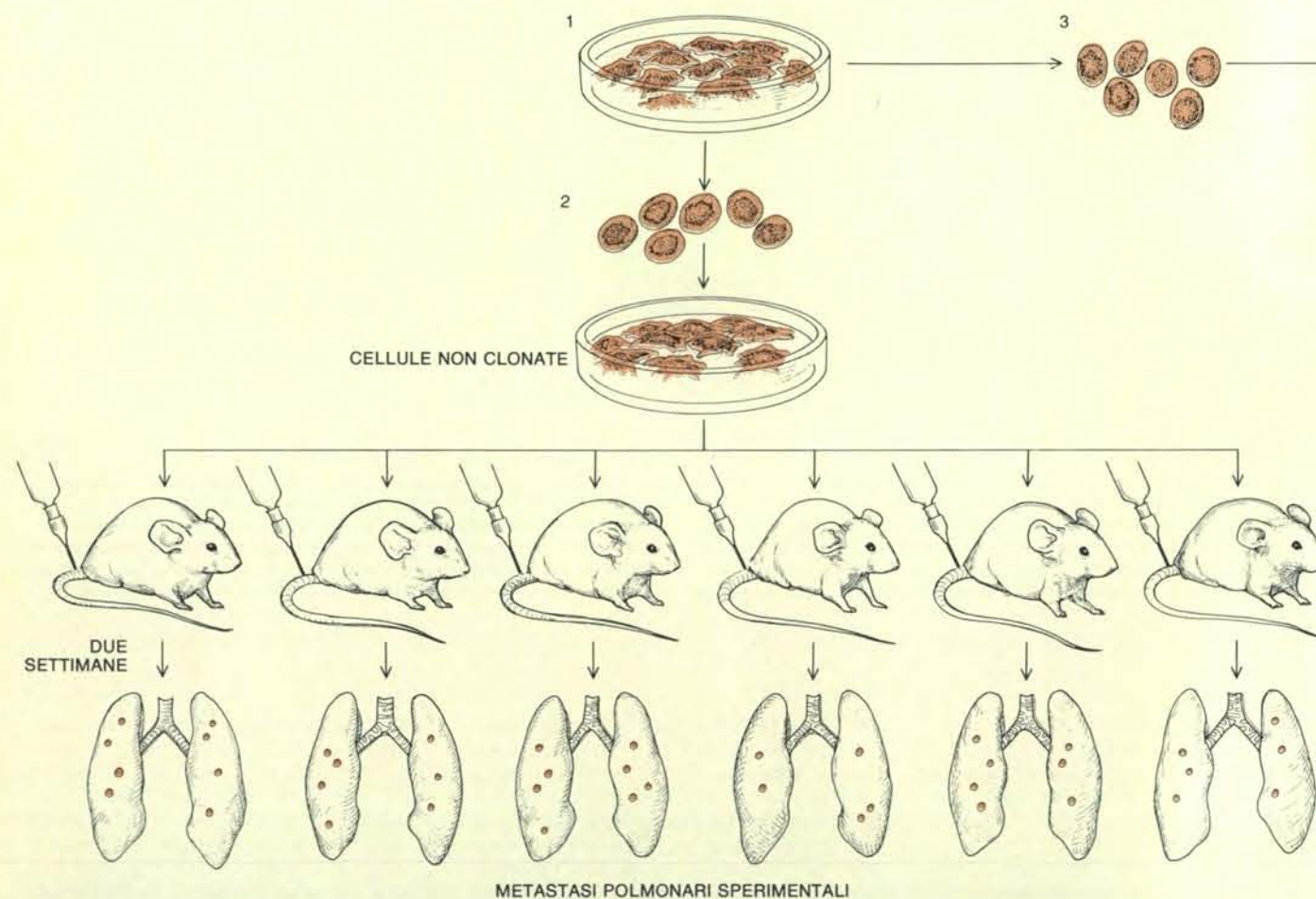
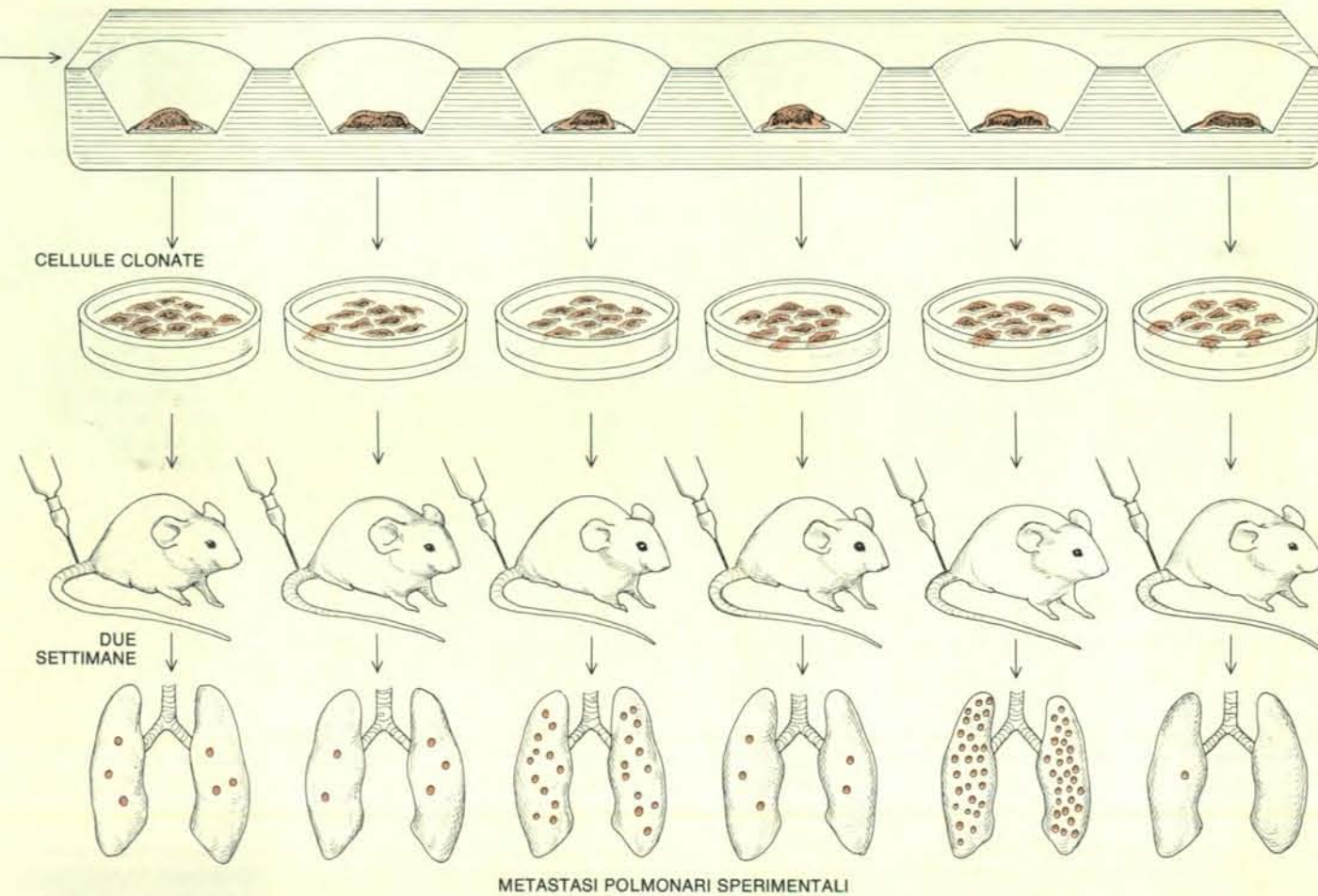
Se il numero di colonie tumorali nei polmoni dei topi che ricevevano le diverse sottopopolazioni clonali si dimostrava simile al numero di colonie dei topi che ricevevano il ceppo non clonato, veniva dimostrato che il ceppo del tumore originario era omogeneo; se differenti sottopopolazioni clonali davano luogo a quantità di colonie ampiamente diverse se paragonate al ceppo non clonato, ciò indicava, invece, che il ceppo originario non selezionato era eterogeneo, con cellule a basso e ad alto potenziale metastatico preesistenti in esso. Quest'ultimo risultato fu ottenuto in tutti e tre gli esperimenti. L'implicazione è che la popolazione di cellule tumorali è formata da elementi altamente eterogenei in rapporto alle proprietà che governano la malignità e che alcune sottopopolazioni altamente metastatiche esistono in un tumore prima che il processo della progressione tumorale e della selezione delle popolazioni cellulari con caratteristiche di aumentata malignità abbia luogo.

Quando una cellula normale si trasforma in una cellula tumorale, la superficie cellulare mostra antigeni nuovi (o per lo meno modificati), ossia strutture molecolari che il sistema immunitario riconosce come estranee. La comparsa di questi antigeni associati al tumore può stimolare nell'ospite una risposta immunitaria che porta alla intercettazione e alla distruzione delle cellule tumorali. F. M. Burnet del Walter and Eliza Hall Institute of Medical Research di Melbourne, ha proposto che le cellule si trasformino periodicamente, ma che siano impedite dal formare tumori da un sistema generalizzato di «sorveglianza immunitaria» e che i pochi tumori che compaiono si originino occasionalmente dalle cellule trasformate che riescono a superare il sistema di sorveglianza. L'efficacia del sistema immunitario nella protezione contro la comparsa e lo sviluppo dei tumori è stata studiata da altri ricercatori. Richmond T. Prehn del Jackson Laboratory e Osias Stutman del Memorial Sloan-Kettering Cancer Center sono tra coloro che mettono in dubbio il complesso della sorveglianza immunitaria come efficace difesa antitumorale.

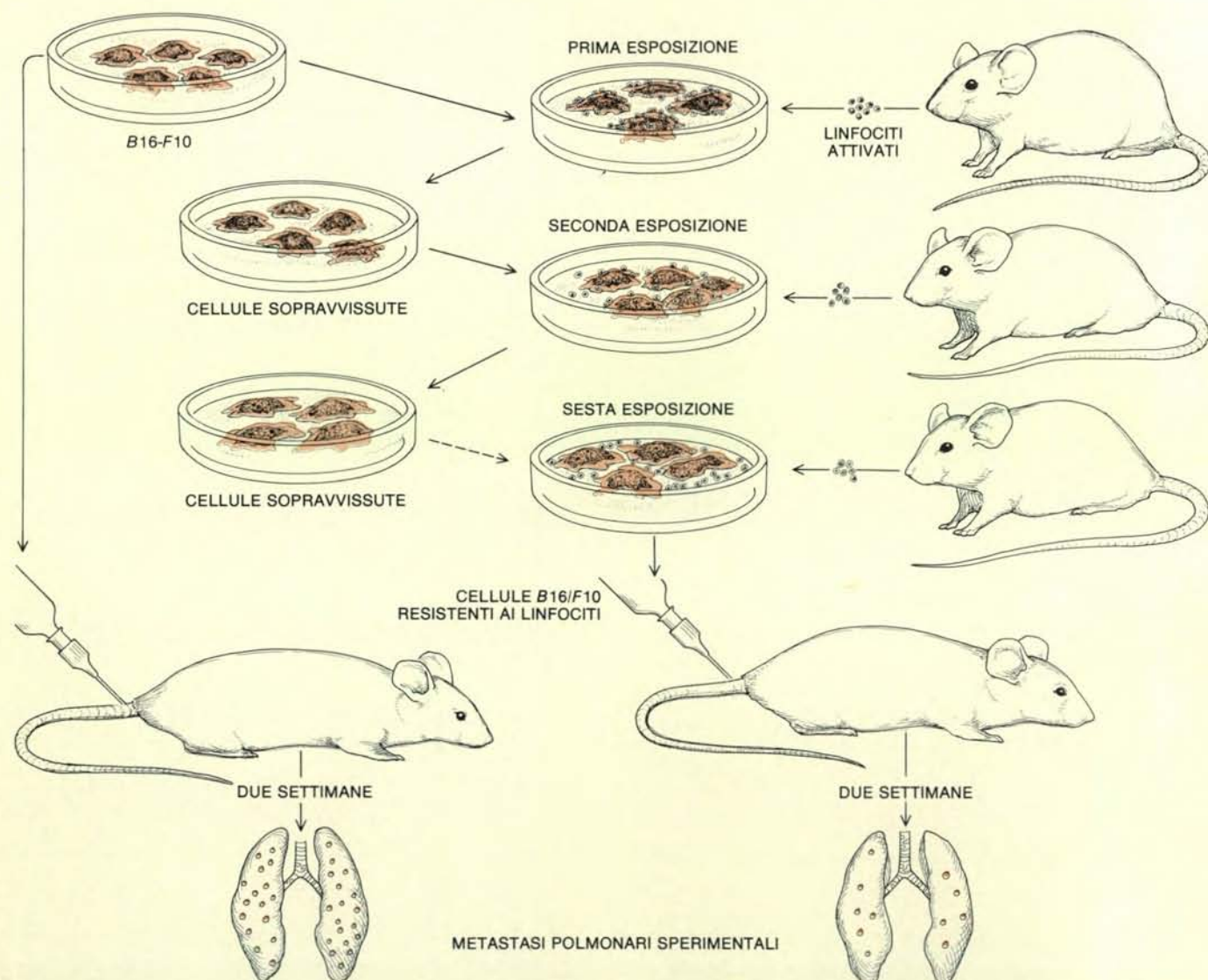
Certamente le risposte immunitarie interferiscono con lo sviluppo e la metastasi di alcuni tumori maligni umani e animali. Alcuni ricercatori, tra cui Robert W. Baldwin della Università di Nottingham, Luka Milas dell'Istituto Rudjer Bosckovic in Iugoslavia, Vilas V. Likhite della Harvard Medical School e John E. Castro della Royal Poatgraduate Medical School di Londra, hanno dimostrato che la metastasi di determinati tumori può essere inibita se il sistema immunitario viene stimolato mediante introduzione di microrganismi quali il bacillo di Calmette-Guérin (BCG) e il *Corynebacterium parvum*. Il meccanismo di questa reazione non è ben compreso: forse essa è resa possibile da uno stimolo non specifico del sistema immunitario o da una risposta specifica agli antigeni presenti sulla superficie del microrganismo che sono simili agli antigeni presenti sulle cellule tumorali.

Il controllo della metastasi non è, tuttavia, sempre correlato con la «forza» del sistema immunitario. Gli esperimenti spesso non sono riusciti a dimostrare che la metastasi di origine ematica sia più frequente negli animali il cui sistema immunitario è stato depresso sperimentalmente che non negli animali con in sistema immunitario intatto. È certo che l'aumento delle difese immunitarie in alcuni pazienti e in alcuni animali da esperimento si traduce in un aumento delle metastasi. Quando i Fisher depressero l'immunità antitumorale dei topi iniettando anticorpi che distrussero la risposta dei linfociti (che costituiscono una componente cellulare importante del sistema immunitario) nei confronti delle cellule tumorali, essi trovarono che si formava un numero minore - piuttosto che maggiore - di metastasi. D'altro canto, Michael Feldman del Weizmann Institute of Science di Israele notò che l'iniezione endovenosa di grandi quantità di linfociti che erano stati attivati per uccidere le cellule del carcinoma era in grado di inibire la metastasi di quel carcinoma negli animali da esperimento.

Tali risultati apparentemente contraddittori spinsero Prehn a suggerire che possono essere coinvolti due sistemi immunitari tra loro competitivi, uno dei quali responsabile dell'eliminazione dei tumori e l'altro responsabile della stimolazione dei tumori. Prehn scoprì che introducendo un piccolo numero di linfociti antitumorali in topi portatori di tumore stimolava la crescita del tumore, mentre l'iniezione di un gran numero di linfociti inibiva la crescita del tumore. Esperimenti analoghi sono stati condotti con il sistema di cellule *B*16 di melanoma. L'iniezione in topi di una piccola quantità di linfociti insieme a cellule *B*16 portava a un incremento significativo delle metastasi polmonari, ma quando una grande quantità di linfociti erano mescolati con cellule tumorali, le metastasi divenivano meno frequenti.

In che modo una piccola quantità di linfociti può stimolare il depositarsi e la sopravvivenza delle cellule di melanoma in circolo? La risposta non è certo definitiva, ma sappiamo che, quando i linfociti e le cellule *B*16 di melanoma sono mescolate in sospensione, esse formano piccoli aggregati. Come ho già detto, la formazione di aggregati incrementa la capacità di deposizione e di sopravvivenza delle cellule tumorali. Ciò potrebbe avvenire semplicemente perché essi si arrestano nei capillari polmonari più rapidamente o perché un gruppo di sostanze chiamate linfochine, che possono essere rilasciate dai linfociti dopo che essi hanno interagito con le cellule tumorali, favoriscono in qualche modo la sopravvivenza di cellule maligne.

L'eterogeneità metastatica di una popolazione di cellule tumorali è stata dimostrata mediante esperimenti eseguiti nel laboratorio di Fidler e dell'autore. In ciascun caso una popolazione di cellule tumorali non selezionate *(1)* venne suddivisa in due frazioni. Una frazione *(2)* venne iniettata per via endovenosa in numerosi gruppi di topi. L'altra frazione *(3)* venne usata per produrre un gran numero di cloni, ciascuno discendente da una singola cellula; ogni clone fu fatto crescere e le cellule così ottenute furono iniettate in altri gruppi di topi. Il numero di metastasi derivate dalle cellule non clonate e da ciascuno dei cloni differiva di molto, indicando che la popolazione originaria *(1)* doveva essere stata altamente eterogenea. Diversamente il numero di metastasi derivate dalle cellule non clonate e da ogni clone sarebbe stato simile.

Perché allora, linfociti in grandi quantità inibirebbero la metastasi? Nell'attività di aggregazione delle cellule *B*16 una grande quantità di linfociti è più efficace che non una piccola, e ciò dovrebbe facilitare presumibilmente la deposizione e l'impianto delle cellule tumorali. D'altra parte, una grande quantità di linfociti inibisce le metastasi uccidendo le cellule maligne. Potrebbe darsi che l'effetto antimetastatico di grandi quantità di linfociti sia dovuto in realtà alla presenza tra di essi di una piccola sottopopolazione di linfociti «killer» che vengono attivati specificatamente per sopprimere le cellule tumorali. O forse nei diversi esperimenti i linfociti sono stati in qualche modo contaminati da alcune cellule immunitarie funzionalmente differenziate le quali possono essere responsabili della soppressione delle cellule tumorali.

Nel tentativo di sapere se i linfociti erano o meno direttamente responsabili della inibizione delle metastasi, Fidler mise in coltura cellule *B*16-*F*10 di melanoma con linfociti attivati. La maggior parte delle cellule tumorali vennero soppresse, ma alcune sopravvissero; queste furono fatte crescere e sottoposte di nuovo all'azione dei linfociti attivati, a loro volta le cellule sopravvissute vennero isolate, coltivate e ricoperte di linfociti attivati. Dopo sei di queste selezioni per la resistenza ai linfociti, le cellule sopravvissute *B*16-*F*10 erano completamente resistenti alla soppressione da parte dei linfociti. Queste cellule resistenti furono iniettate per via endovenosa nei topi. Esse formarono un numero di metastasi minore di quelle delle cellule *F*10 non trattate, non dimostrandosi, di fatto, più attive delle cellule *B*16-*F*1 a bassa capacità di metastatizzazione. Inoltre, esse non riuscivano a formare ammassi se mescolate con linfociti in sospensione. Fidler concluse che l'effetto normale dei linfociti sulle cellule *B*16 è di farle aggregare quando sono in circolo e quindi di facilitare la metastasi e che l'inibizione della metastasi osservata con alti rapporti di linfociti per cellula tumorale è probabilmente causata da qualche altro tipo di cellula immune che è presente con i linfociti.

Questa ipotesi è stata convalidata da esperimenti con topi «nudi», un ceppo di topi privi di pelo che mancano della forma matura del timo e quindi non posseggono immunità: per esempio, essi accettano trapianti di pelle da altre specie. Quando iniettammo cellule *B*16 marcate in topi nudi, le cellule si depositarono e sopravvissero in quantità minori di quanto non facessero in topi dotati di immunità normale. Un'iniezione di linfociti ottenuti da

L'effetto dei linfociti sulla metastasi è stato studiato da Fidler. Egli espose cellule *B*16-*F*10 di melanoma all'azione di linfociti attivati, vale a dire, a linfociti estratti dalla milza e dai linfonodi di topi immuni nei confronti di cellule tumorali *B*16. Alcune cellule *B*16-*F*10 vennero soppresse; le sopravvissute furono nuovamente sottoposte all'azione dei linfociti attivati. Dopo la ripetizione per sei volte del procedimento, le rimanenti cellule *B*16-*F*10 erano diventate del tutto resistenti all'attacco dei linfociti. Tuttavia quando le cellule resistenti furono iniettate in topi, davano origine a metastasi polmonari in numero molto inferiore alle cellule *B*16-*F*10 non trattate. La resistenza ai linfociti non aumentava il potenziale metastatico delle cellule, ma anzi lo riduceva eliminando le fondamentali interazioni tra il linfocita e la cellula tumorale.

topi normali in topi nudi un giorno prima dell'iniezione di cellule tumorali ripristinava in modo parziale lo schema normale della metastasi. Se ne deduce ancora che, almeno nel sistema *B*16, i linfociti stimolano la metastasi sperimentale.

È vero che in altri esperimenti, tra cui alcuni condotti nei laboratori di Feldaman e di Baldwin presso il Chester Beatty Research Institute, le risposte immunitarie mediate dai linfociti sono risultate efficaci nel prevenire la diffusione metastatica o nell'eliminare micrometastasi già in corso. Il fatto che non si possa prevedere con facilità se l'immunità dell'ospite incrementi o inibisca la diffusione del tumore, significa che l'immunità e la metastasi non stanno in un rapporto semplice. I risultati sperimentali sono avvalorati da diverse osservazioni cliniche: procedure progettate per rafforzare le difese immunitarie di un paziente affetto da tumore maligno hanno inibito la metastasi in alcuni casi e stimolato la metastasi in altri. La chemioterapia o la immunoterapia per il cancro possono, oltre a sopprimere alcune cellule tumorali, portare alla selezione di sottopopolazioni di cellule resistenti, altamente metastatiche. Il problema di individuare una terapia che non dia luogo a indesiderati effetti selettivi, costituisce materia di dibattito in ogni singolo caso.

Certamente lo studio di terapie efficaci dovrebbe tenere conto del fatto che solo alcune cellule altamente metastatiche in una popolazione tumorale possono in ultima analisi arrecare il danno maggiore. Esperimenti recenti condotti nel mio laboratorio da Reuben Lotan dimostrano che ciascuna cellula di melanoma *B*16 selezionata o il clone *B*16 non selezionato rispondono in modo diverso ai farmaci somministrati durante la chemioterapia. L'eterogeneità delle popolazioni tumorali può essere responsabile di alcuni insuccessi clinici e sperimentali della chemioterapia se le sottopopolazioni resistenti sopravvivono al trattamento farmacologico. Un farmaco può essere prescelto perché inibisce la crescita del tumore primitivo o della sua controparte colturale e tuttavia può non essere in grado di arrestare la crescita di alcune sottopopolazioni cellulari del tumore che sono presenti nelle micrometastasi.

I tentativi futuri dovranno senza dubbio puntare sulla comprensione della biologia dello sviluppo tumorale e del processo metastatico. Studi di ceppi cellulari tumorali che metastatizzano preferenzialmente in organi particolari, possono svelare il meccanismo di controllo della deposizione di queste cellule e del loro impianto e crescita in un determinato ambiente; studi di ceppi cellulari che sono resistenti a determinati farmaci o al sistema immunitario dovrebbero permettere di comprendere perché alcuni tumori resistono a tutti i tentativi terapeutici. Una maggiore conoscenza sul processo di metastatizzazione e sullo sviluppo dei tumori negli animali dovrebbero portare allo studio di terapie più razionali ed efficaci per il trattamento dei tumori maligni.

Lo spettro dell'idrogeno atomico

Per quasi un secolo la luce emessa dall'atomo più semplice ha fornito una base sperimentale alle teorie della struttura della materia. Oggi le ricerche sullo spettro dell'idrogeno continuano, facilitate dai laser

di Theodor W. Hänsch, Arthur L. Schawlow e George W. Series

Lo spettro dell'atomo di idrogeno si è rivelato la stele di Rosetta della fisica moderna: una volta decifrato questo insieme di righe, si potrebbero capire molte altre cose. Da rilevare il fatto che è stato lo sforzo di spiegare lo spettro della luce emessa dall'atomo di idrogeno a ispirare ampiamente le leggi della meccanica quantistica. Tali leggi sono risultate valide non solo per l'atomo di idrogeno ma anche per altri atomi, per le molecole e la materia condensata, e costituiscono la base fondamentale della chimica moderna, della fisica dello stato solido e di scienze applicate quale l'elettronica.

La posizione centrale dell'atomo di idrogeno nella storia della fisica del secolo XX potrebbe far credere che lo spettro sia conosciuto da tempo in tutti i suoi particolari, mentre non è così. Soltanto negli ultimi anni sono stati chiariti alcuni degli aspetti più delicati dello spettro e molti altri non sono ancora stati osservati direttamente. La misurazione delle posizioni delle righe spettrali rimane ancora oggi un significativo banco di prova delle previsioni della teoria dei quanti.

I progressi più recenti nell'analisi dello spettro si possono attribuire a un nuovo strumento della spettroscopia: il laser. Infatti, a causa della struttura fine dello spettro, i dettagli più fini si possono risolvere soltanto con luce altamente monocromatica o limitata a un ristretto intervallo di lunghezze d'onda. Il laser è una sorgente di luce di questo tipo e per merito di alcune ingegnose tecniche di laboratorio è diventato uno strumento spettroscopico con risoluzione mai raggiunta in precedenza.

Lo spettro a righe

Quando la luce emessa da un filamento arroventato subisce la dispersione in lunghezza d'onda da parte di un prisma o di un reticolo di diffrazione, il risultato è un ventaglio continuo di colori, ma lo spettro di un gas puro rarefatto di atomi o molecole è formato da righe discrete. Se si registra lo spettro della luce emessa dal gas, esso appare come una successione di righe chiare su uno sfondo scuro. Se la luce attraversa il gas, vengono assorbite lunghezze d'onda discrete, originando righe scure su uno sfondo chiaro.

L'idrogeno è il più semplice degli atomi, essendo costituito da un solo elettrone e da un nucleo formato a sua volta da un solo protone, e quindi ci si può aspettare che abbia lo spettro più semplice. Lo spettro non è però di agevole registrazione. La riga più intensa è stata rivelata nel 1853 da Anders Jonas Ångström. (Per molto tempo l'unità di misura più usata per la lunghezza d'onda della luce è stata - dal suo nome - l'angstrom, pari a 10^{-8} centimetri.) Nei due decenni successivi furono osservate altre tre righe, ma la prima serie estesa di righe dell'idrogeno atomico non è stata scoperta in laboratorio ma nello spettro di stelle. Nel 1881, lavorando con le prime fotografie di spettri stellari, Sir William Huggins identificò 10 righe come emissioni dell'idrogeno atomico.

Può apparire sorprendente che le righe dello spettro dell'idrogeno siano state viste in osservazioni astronomiche prima di essere osservate in esperimenti terrestri. Le difficoltà che si incontrano nel misurare lo spettro in laboratorio non stanno nel rivelare le righe, ma nel preparare idrogeno atomico puro. L'idrogeno gassoso ordinario è formato da molecole biatomiche (H_2), che hanno uno spettro molto più complesso dello spettro dell'atomo di idrogeno isolato. La scissione delle molecole richiede più energia di quanta se ne possa fornire con la maggior parte delle eccitazioni termiche, quali una fiamma. Un dispositivo che fornisce l'energia in modo molto efficace è il tubo a scarica nei gas, nel quale gli elettroni accelerati dall'applicazione di una tensione dissociano le molecole. Gradualmente, si sono scoperti mezzi per evidenziare lo spettro atomico a spese di quello molecolare.

Le righe spettrali rivelate da Huggins sono comprese in un intervallo di lunghezza d'onda che si estende dalla regione rossa dello spettro visibile all'ultravioletto vicino. Le prime due righe sono molto distanti, ma le successive appaiono a intervalli decrescenti con regolarità e quelle alle minori lunghezze d'onda sono strettamente raggruppate. Nel 1885, lavorando su misure astronomiche, Johann Jakob Balmer scoprì di poter spiegare le posizioni di tutte le righe conosciute applicando una semplice formula empirica. L'intero insieme di righe è da allora noto come serie di Balmer. Un altro gruppo di righe, la serie di Lyman, si trova nell'ultravioletto lontano e vi sono altre serie a lunghezze d'onda maggiori. Le singole righe all'interno di ogni serie sono indicate con lettere greche, partendo dalla riga di lunghezza d'onda maggiore. Così la riga rossa luminosa vista per la prima volta da Ångström è la riga Balmer-alfa, la successiva è la riga Balmer-beta e così via. Per la sua intensità la riga Balmer-alfa è chiamata riga idrogeno-alfa, o H-alfa.

Nel 1889 Johannes Rydberg scoprì che gli spettri a righe di molti elementi, in particolare quelli dei metalli alcalini, si potrebbero spiegare con una sola formula empirica. Una successiva versione della formula, equivalente all'equazione di Rydberg ma più esplicita, ha la forma

$$\frac{1}{\lambda} = R\left[\frac{1}{(m+b)^2} - \frac{1}{(n+c)^2}\right].$$

dove (la lettera greca lambda) è la lunghezza d'onda di una particolare riga in uno spettro atomico, m ed n sono numeri che assumono valori interi successivi (1, 2, 3 e così via) e R, b e c sono costanti. I valori di b e di c dipendono da quale serie di righe si sta misurando, mentre R è la stessa per qualsiasi riga di qualsiasi elemento. R si chiama oggi costante di Rydberg: se la lunghezza d'onda si esprime in metri, R vale $1{,}097 \times 10^7$ e le sue dimensioni sono quelle dell'inverso di una lunghezza. Dopo aver verificato la validità della sua formula per un certo numero di spettri, Rydberg scrisse: «Avevo appena finito di provare varie forme della funzione, quando ho sentito la comunicazione di Mr. Balmer sulle righe spettrali dell'idrogeno atomico. Fui felice nello scoprire che la sua formula è un caso particolare della mia con lo stesso valore di R e con $c = 0$.» Quindi per l'atomo di idrogeno l'equazione assume una forma più semplice:

$$\frac{1}{\lambda} = R\left[\frac{1}{m^2} - \frac{1}{n^2}\right].$$

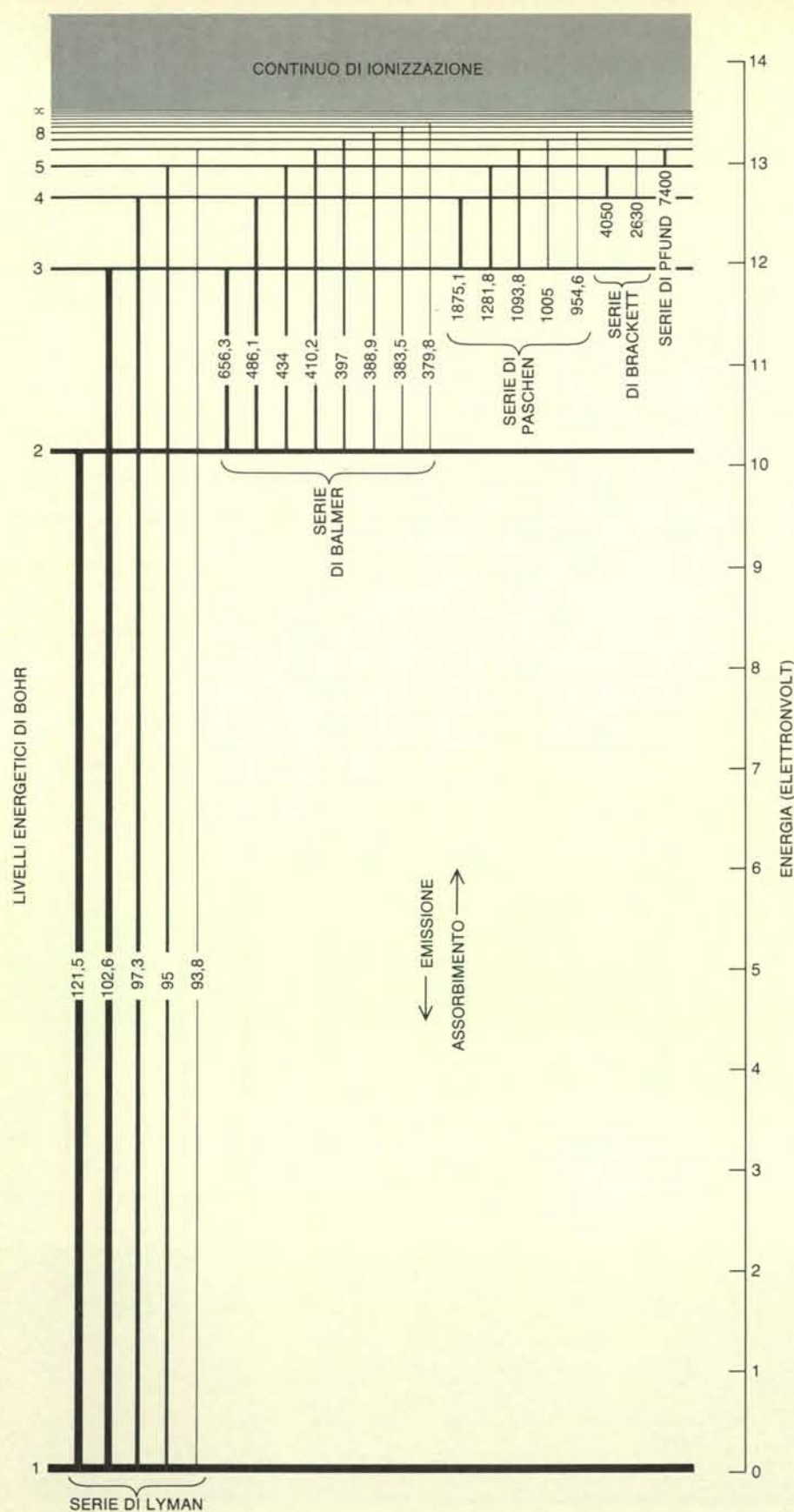
Quando in questa formula si sostituiscono gli opportuni valori di m e di n, si ottengono le lunghezze d'onda di tutte le righe dello spettro dell'idrogeno.

La struttura dell'atomo

Nel 1912 Niels Bohr trascorse la maggior parte dell'anno all'Università di Manchester lavorando nel laboratorio di Ernest Rutherford, che già aveva dato un vitale contributo alla comprensione della struttura atomica. Già allora le formule di Balmer e di Rydberg erano ben note agli spettroscopisti ed erano state verificate a fondo, ma Bohr non aveva mai sentito parlare di nessuna delle due. Quando gli parlarono dell'equazione di Balmer, egli fu in grado di ricavarla quasi subito dalle proprietà dell'atomo di idrogeno. Abraham Pais e T. D. Lee raccontano che molti anni dopo a Bohr venne chiesto come mai non conoscesse la formula di Balmer. Egli spiegò che in quei tempi la maggior parte dei fisici considerava gli spettri atomici talmente complessi da far dubitare che fossero una parte della fisica fondamentale. Si credeva, egli disse, che fossero come i suoni di un pianoforte, che dipendono dalla struttura dello strumento nel suo complesso.

Il riferimento alle note musicali non era banale. Nel secolo XIX molti ricercatori avevano avanzato l'ipotesi che la luce emessa dagli atomi si potesse spiegare con qualche analogia con i modi di vibrazione di un solido. La presenza di righe multiple in ogni serie si poteva così spiegare supponendo che la prima riga fosse il modo fondamentale e le altre sovratoni, o armoniche. Dal momento, però, che nessuno spettro ottico si poteva risolvere in una frequenza fondamentale e nei suoi sovratoni, si intrapresero ricerche più dirette sull'atomo stesso.

Il contributo di Rutherford dimostrava che l'atomo non è un corpo solido, ma è invece costituito da un piccolo nucleo denso carico positivamente e da un opportuno numero di elettroni carichi nega-

Quando la luce emessa da atomi di idrogeno viene dispersa secondo la lunghezza d'onda si forma uno schema di righe distinte. Le righe sono quelle della serie di Balmer e le loro lunghezze d'onda approssimate sono espresse in nanometri (un nanometro è uguale a 10^{-9} centimetri). La fotografia a colori è stata eseguita da Jon Brenneis e da uno degli autori (Schawlow) nel laboratorio di quest'ultimo all'Università di Stanford. Lo spettro in bianco e nero venne registrato nel 1927 da Gerald Herzberg, che era allora alla Technische Universität di Darmstadt in Germania. Gli intervalli tra le righe sono diversi nelle due fotografie perché Brenneis e Schawlow, per disperdere la luce, usarono un reticolo di diffrazione, mentre Herzberg ricorse a un prisma; il prisma introduce una distorsione, dato che l'indice di rifrazione del vetro non è una funzione lineare della lunghezza d'onda. In entrambi i casi la sorgente dello spettro era un tubo a scarica nei gas nel quale le molecole di idrogeno (H_2) vengono dissociate in atomi da un potenziale elettrico. Nella fotografia a colori le bande gialla e verde e il fondo continuo blu sulla destra non fanno parte dello spettro dell'idrogeno atomico, ma sono dovute a contaminazioni molecolari nel tubo di scarica. Herzberg eliminò tali contaminazioni raffreddando il tubo in azoto liquido. Le righe a destra della riga Balmer-delta, in blu chiaro nella fotografia a colori, appartengono in realtà alla regione ultravioletta dello spettro.

Le transizioni tra gli stati dell'atomo di idrogeno danno origine allo spettro a righe. Ogni stato è designato da un numero intero, chiamato numero quantico principale del corrispondente livello energetico di Bohr. Quando un atomo passa da uno stato all'altro, la differenza di energia appare come un quanto di radiazione, la cui energia è proporzionale direttamente alla frequenza della radiazione e inversamente alla lunghezza d'onda. L'assorbimento di radiazione stimola una transizione a uno stato di maggiore energia; un atomo che passa a uno stato di minore energia emette radiazione. Lo spettro comprende varie serie di righe con un livello inferiore in comune. Le lunghezze d'onda sono espresse in nanometri, l'intensità relativa delle righe è indicata dallo spessore.

tivamente per formare un sistema stabile ed elettricamente neutro. Bohr cercò di spiegare gli spettri atomici nel contesto di questo modello. Nel fare ciò egli rischiò di apparire ridicolo ai suoi contemporanei, ipotizzando che all'interno dell'atomo le leggi della fisica acquisite non sono valide. Tali leggi prevedevano che un qualsiasi elettrone legato avrebbe dovuto irradiare tutta la sua energia e cadere nel nucleo. Bohr propose l'esistenza in un atomo di uno stato in cui l'energia di un elettrone non veniva dissipata, uno stato stazionario. Egli suggerì che la luce potesse essere emessa o assorbita soltanto quando un elettrone passa da uno stato stazionario a un altro. La frequenza della luce è determinata dalla differenza di energia tra i due stati, secondo il principio quantistico introdotto da Max Planck, e si trova dividendo la differenza di energia per la costante di Planck, indicata con h.

Nel modello di Bohr le formule di Balmer e di Rydberg possiedono un'evidente interpretazione fisica. L'espressione $(1/m^2 - 1/n^2)$ è proporzionale alla differenza di energia tra due stati dell'atomo: gli interi m ed n individuano gli stati stessi. Bohr fece però di più che limitarsi a spiegare questa relazione empirica; egli giunse a calcolare la costante di Rydberg in funzione di tre grandezze fondamentali, la carica elettrica dell'elettrone, e, la massa dell'elettrone, m, e la costante di Planck, h. Se, per amor di semplicità, si attribuisce al nucleo atomico una massa infinita, la costante di Rydberg è data dall'equazione

$$R = \frac{2\pi^2 m e^4}{h^2} .$$

Ulteriori ritocchi hanno reso più complessa la formula empirica di Rydberg per le lunghezze d'onda delle righe spettrali e così la costante di Rydberg è oggi definita come tale combinazione di m, e ed h.

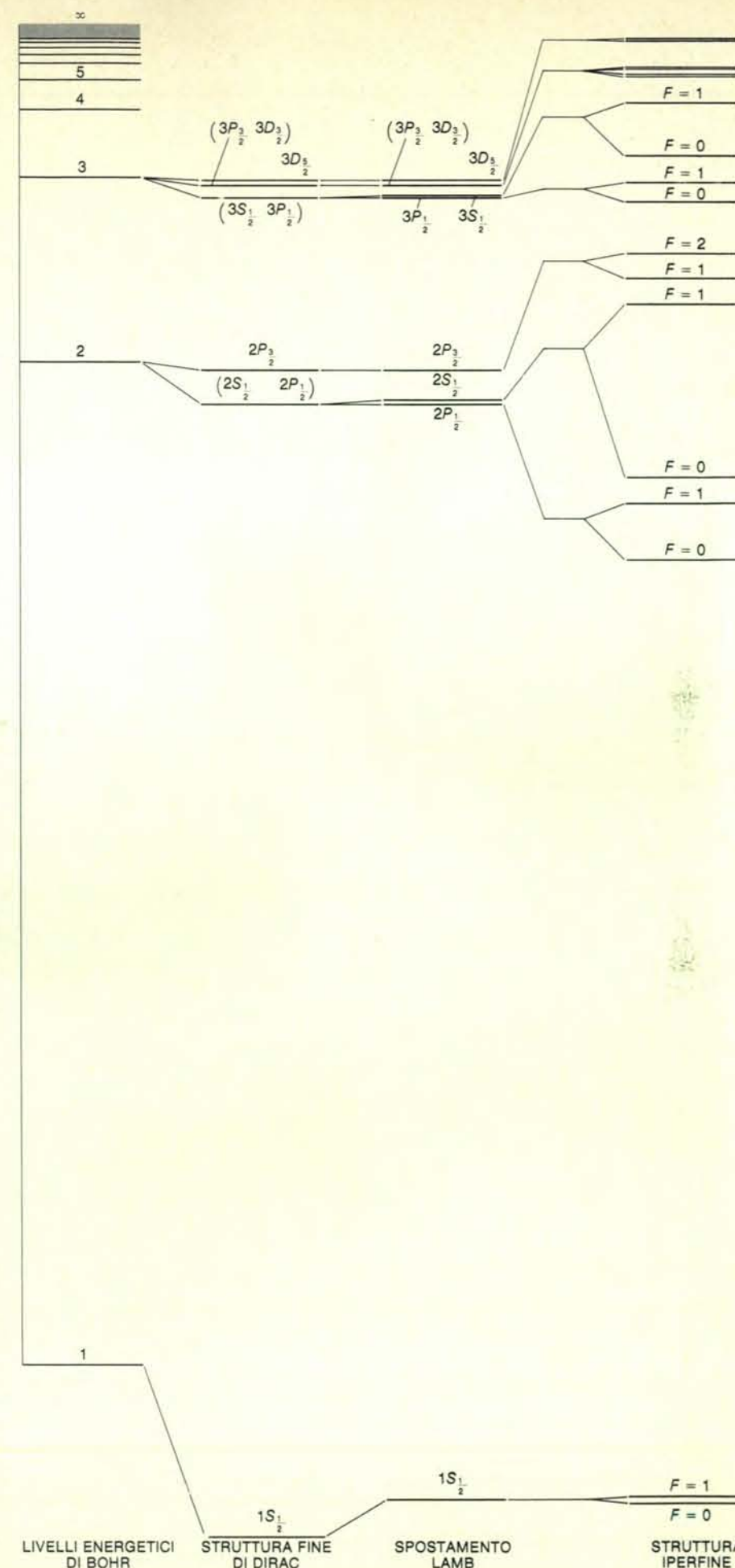
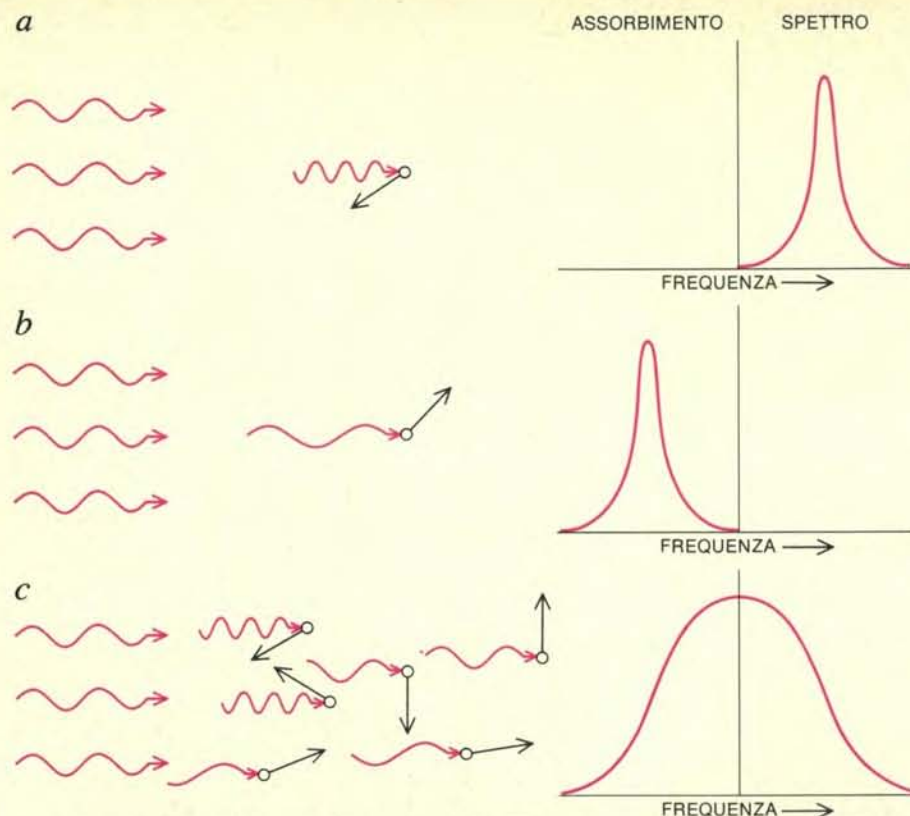
Bohr immaginò che gli elettroni in un atomo potessero percorrere orbite circolari. Nel calcolare l'energia degli stati stazionari egli cominciò a supporre che per orbite di diametro molto grande la frequenza della luce emessa corrispondesse alla frequenza dell'elettrone orbitante, previsione questa della fisica classica. Questo «principio di corrispondenza» condusse alla conseguente conclusione che per uno stato stazionario indicato con l'intero n il momento angolare dell'elettrone orbitante è uguale a $nh/2\pi$. Quindi anche il momento angolare, come l'energia, può variare soltanto di quantità discrete. Questa «quantizzazione» del momento angolare è di significato fondamentale: senza di essa l'atomo potrebbe irradiare a qualsiasi frequenza e lo stato non sarebbe più stazionario.

La meccanica quantistica

L'atomo di Bohr costituì un grande progresso concettuale sulle precedenti teorie, ma si rivelò ben presto inadeguato per spiegare le proprietà osservate negli spettri atomici, perfino quelle dello spettro più semplice, quello dell'idrogeno. Infatti, un dettaglio dello spettro dell'idrogeno che la teoria di Bohr non era in grado di spiegare era noto da 20 anni quando venne proposto il modello. Nel 1892 A.A. Michelson aveva impiegato l'interferometro di sua invenzione per esaminare la forma delle singole righe spettrali, cioè la variazione di intensità in funzione della lunghezza d'onda. Egli aveva scoperto che la riga Balmer-alfa non è affatto singola, ma ha due componenti le cui lunghezze d'onda differiscono di 0,014 nanometri.

Bohr avanzò l'ipotesi che la scissione della riga Balmer-alfa si poteva spiegare se l'orbita dell'elettrone nell'atomo di idrogeno fosse stata non circolare ma ellittica. Una correzione al moto orbitale imposta dalla teoria della relatività ristretta avrebbe in tal caso scisso ogni stato stazionario in un gruppo di stati. Una più completa esposizione di questo concetto fu successivamente presentata da Arnold Sommerfeld, che ricavò le scissioni osservate dal calcolo delle proprietà delle orbite. Nei suoi calcoli egli introdusse un importante numero adimensionale, chiamato costante di struttura fine, uguale a $2\pi e^2/hc$. Il valore numerico di questa costante, circa 1/137, è sempre stato per i fisici una fonte di suggestive ipotesi.

Una trattazione più soddisfacente della scissione delle righe non fu possibile fino all'introduzione della meccanica quantistica negli anni venti. Furono ricavate due formulazioni della meccanica quantistica, una da Erwin Schrödinger, l'altra da Werner Heisenberg e Max Born: si scoprì ben presto che esse erano matematicamente equivalenti. Un concetto essenziale in entrambe le formulazioni era che non si potesse definire esattamente il moto di un elettrone, ma lo si potesse solo descrivere in termini di probabilità. Entrambe le teorie prevedevano i livelli energetici di Bohr per l'atomo di idrogeno, e quindi le stesse lunghezze d'onda delle righe spettrali, e le previsioni si potevano estendere a sistemi atomici più complessi. Ancora una volta per poter spiegare la scissione delle righe dell'idrogeno si dovevano apportare alla teoria correzioni di natura relativistica.

La scissione dei livelli energetici è dovuta principalmente alle interazioni relativistiche e magnetiche correlate ai momenti angolari degli atomi. L'elettrone singolo di un atomo di idrogeno può avere momento angolare sia orbitale che di spin e le varie combinazioni possibili di queste grandezze hanno di solito energie differenti. Così i livelli energetici di Bohr vengono scissi in componenti chiamati struttura fine dei livelli. Un lieve effetto chiamato spostamento Lamb sposta alcune componenti a struttura fine, creando ulteriori scissioni. Infine, l'interazione del momento magnetico dell'elettrone con quello del nucleo dà origine a una struttura iperfine. La struttura fine e lo spostamento Lamb sono qui indicati in scala maggiore rispetto ai livelli di Bohr, mentre le scissioni iperfini del secondo e terzo livello di Bohr sono indicate con ingrandimento ancora maggiore.

L'allargamento Doppler nasconde molti particolari dello spettro dell'atomo di idrogeno allargando ogni componente stretta in una banda di lunghezze d'onda. Dato che gli atomi di un gas hanno velocità casuali, in un istante qualsiasi alcuni si muovono verso una sorgente luminosa, mentre altri si allontanano da essa. Per un atomo che si sta avvicinando alla sorgente *(a)* la luce appare con frequenza maggiore (o lunghezza d'onda minore) che nel sistema di riferimento del laboratorio. Per un atomo che si allontana *(b)* la frequenza appare minore. A causa di questi effetti Doppler, la frequenza misurata di una componente spettrale è diversa per atomi dotati di differenti velocità. Il profilo di assorbimento per tutti gli atomi del gas *(c)* non è un picco netto ma una curva larga.

Quando però, alla fine degli anni venti, furono disponibili misure più precise della forma delle righe dello spettro dell'idrogeno, apparve chiaro che si rendeva necessario un ulteriore ritocco alla teoria. Il ritocco venne apportato quando si scoprì che l'elettrone poteva acquistare un momento angolare non soltanto orbitando attorno al nucleo ma anche ruotando attorno al proprio asse. Il concetto di spin dell'elettrone fu introdotto da George Uhlenbeck e Samuel A. Goudsmit per spiegare alcune misure di spettri degli atomi di metalli alcalini. Introducendo nella teoria lo spin dell'elettrone fu possibile prevedere con ragionevole precisione non soltanto le lunghezze d'onda delle righe spettrali ma anche le loro intensità, cosa che non era possibile con la teoria di Bohr-Sommerfeld. Nel 1928 il concetto di spin dell'elettrone - e anche del protone e del neutrone - ricevette una base più solida quando P. A. M. Dirac scoprì un nuovo modo di costruire una teoria della meccanica quantistica, anch'esso coerente con la teoria della relatività ristretta. Nella teoria di Dirac lo spin dell'elettrone appare come una conseguenza naturale delle equazioni fondamentali piuttosto che come un postulato particolare. Il modulo del momento angolare di spin è $1/2\ h/2\pi$.

La struttura fine

La meccanica quantistica relativistica di Dirac prevede la struttura dello spettro dell'idrogeno con grandi dettagli. Come nelle precedenti teorie ogni livello energetico principale è indicato con un intero n, chiamato numero quantico principale. Per il livello di minima energia (lo stato fondamentale di un atomo) n è uguale a 1, per il primo stato eccitato n è uguale a 2 e così via. Questi interi sono gli stessi che compaiono nella formula delle righe spettrali di Rydberg.

Le righe principali dello spettro - ignorando per il momento qualsiasi scissione in componenti più fini - sono prodotte da transizioni tra stati che hanno numeri quantici principali differenti. La riga Balmer-alfa, per esempio, è dovuta a una transizione tra gli stati $n=2$ e $n=3$. Se un atomo di idrogeno passa da $n=3$ a $n=2$, esso emette luce alla lunghezza d'onda della riga Balmer-alfa; se poi l'atomo assorbe un quanto di luce alla stessa lunghezza d'onda, compie* la transizione opposta, da $n=2$ a $n=3$. Tutte le righe di una serie hanno in comune uno stato di minima energia. Le transizioni che danno origine alla serie di Balmer collegano stati con numeri quantici principali 2 e 3, 2 e 4, 2 e 5, e così via. Nella serie di Lyman tutte le transizioni cominciano dallo stato $n=1$.

Le righe principali dello spettro dell'idrogeno riflettono l'interazione predominante nell'atomo di idrogeno: l'attrazione elettrostatica tra il nucleo e l'elettrone. La scissione di ogni riga riflette interazioni secondarie, effetti relativistici ed effetti magnetici. Gli effetti magnetici compaiono perché l'elettrone rotante, comportandosi come un magnete, è soggetto a un campo magnetico attribuibile al moto efficace del nucleo carico attorno all'elettrone. Questa scissione è chiamata struttura fine dello spettro. Dal momento che anche lo stesso nucleo ha uno spin e si comporta come un magnete, interagisce col campo magnetico derivante dal moto della carica dell'elettrone e anche dalla vicinanza del campo magnetico dell'elettrone. Tali interazioni magnetiche avvengono in scala ancora più piccola e le corrispondenti scissioni delle righe spettrali sono chiamate struttura iperfine.

Per i livelli energetici della struttura fine e iperfine è stato messo a un punto un metodo di classificazione che precisa i moduli dei vari momenti angolari (orbitale, di spin dell'elettrone, di spin del nucleo e totale) in uno stato stazionario. Il numero quantico principale, che corrisponde al livello energetico di Bohr, è dato da uno degli interi della successione 1, 2, 3.... A questo numero viene aggiunta una lettera, come S, P, D o F, che si riferisce al momento angolare orbitale; le lettere stanno al posto degli interi della serie 0, 1, 2, 3,... Un terzo simbolo, che per l'atomo di idrogeno è sempre una frazione come 1/2, 3/2 o 5/2, è collegato alla somma dei momenti angolari orbitale e di spin. Se l'analisi è sufficientemente approfondita da rivelare la struttura iperfine (una interazione che dipende dalle proprietà magnetiche del nucleo), risulta necessario anche un quarto simbolo rappresentato con F, che ha a che vedere con la somma dei momenti orbitale, di spin dell'elettrone e di spin del nucleo. Solitamente stati stazionari che differiscono per uno qualsiasi dei simboli hanno energie differenti, anche se non è sempre così.

La struttura fine scinde il livello $n=2$ dell'atomo di idrogeno in tre sottolivelli e lo stato $n=3$ in cinque sottolivelli. Un semplice calcolo delle possibili combinazioni di questi livelli implica la possibile esistenza di 15 componenti della riga Balmer-alfa. In realtà molte di queste transizioni sono vietate dai principi di conservazione e la teoria di Dirac prevede che la riga debba avere sette componenti. All'epoca in cui Dirac formulò la sua teoria i ricercatori potevano chiaramente riconoscere soltanto due componenti, anche se nel 1925 G. Hansen rivelò una terza componente sotto forma di un allargamento a lato di una delle altre.

Il fattore che limitava la risoluzione delle righe era (ed è tuttora) l'allargamento Doppler causato dai moti di agitazione termica degli atomi. Per l'effetto Doppler la luce emessa da un atomo appare spostata verso una frequenza maggiore se l'atomo si sta avvicinando all'osservatore e verso una frequenza minore se l'atomo si

sta allontanando. Gli atomi in una sorgente luminosa ordinaria si muovono in tutte le direzioni e quindi la luce emessa subisce un aumento di lunghezza d'onda nel rapporto v/c, dove v è la velocità media degli atomi e c è la velocità della luce. L'allargamento è eccezionalmente intenso per l'idrogeno, i cui atomi, più leggeri di qualsiasi altro, si muovono più velocemente. A temperatura ambiente la larghezza di riga addizionale è di circa 12 parti per milione. La velocità degli atomi e quindi la larghezza di riga sono proporzionali alla radice quadrata della temperatura assoluta. La larghezza di riga si può ridurre di un fattore tre o quattro raffreddando la lampada a scarica in idrogeno dalla temperatura ambiente (circa 300 kelvin) alla temperatura dell'idrogeno liquido (20 kelvin).

Lo spostamento Lamb

Negli anni trenta molti ricercatori tentarono di risolvere in modo più completo la struttura fine dell'idrogeno in modo da poter verificare le previsioni della teoria di Dirac. Già allora era possibile analizzare lo spettro del deuterio, l'isotopo dell'idrogeno con massa doppia di quella dell'idrogeno ordinario. La larghezza di una riga spettrale varia in ragione inversa alla radice quadrata della massa atomica e quindi viene ridotta nello spettro del deuterio di un fattore pari a circa 1,4. Anche in questo modo non si riuscì a risolvere alcuna componente addizionale della riga Balmer-alfa e la terza componente non venne ancora completamente separata dalla sua vicina più intensa.

Nel corso di queste ricerche, tuttavia, alcuni osservatori riferirono di alcune discrepanze tra gli spettri misurati e le previsoni della teoria di Dirac. Di particolare interesse era la coppia di livelli a struttura fine $2S_{1/2}$ e $2P_{1/2}$. Questi livelli differiscono nei momenti angolari orbitali, quindi nella distribuzione geometrica media dell'elettrone, ma hanno lo stesso momento angolare totale e la teoria di Dirac prevede che essi debbano avere la stessa energia. In tal modo, pur essendo stati distinti, non dovrebbero dare origine ad alcuna scissione nello spettro. Simon Pasternack interpretò gli esperimenti eseguiti da Robley C. Williams e da William V. Houston e collaboratori dimostrando che i livelli $2S_{1/2}$ e $2P_{1/2}$ non coincidono esattamente. Altri, però, non erano d'accordo.

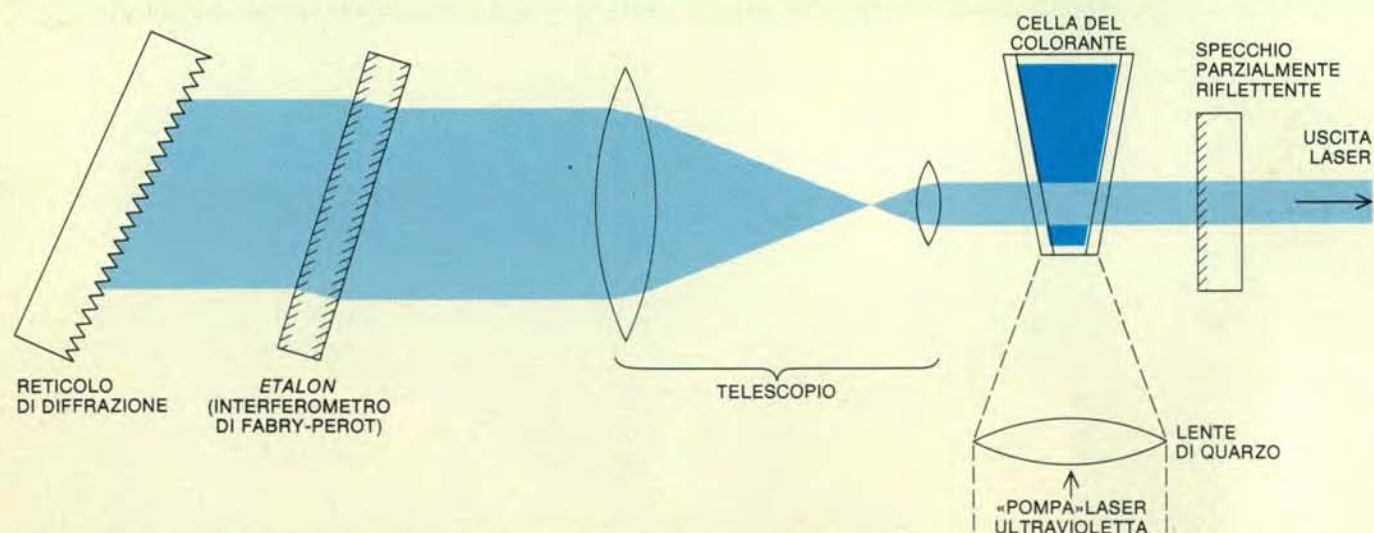
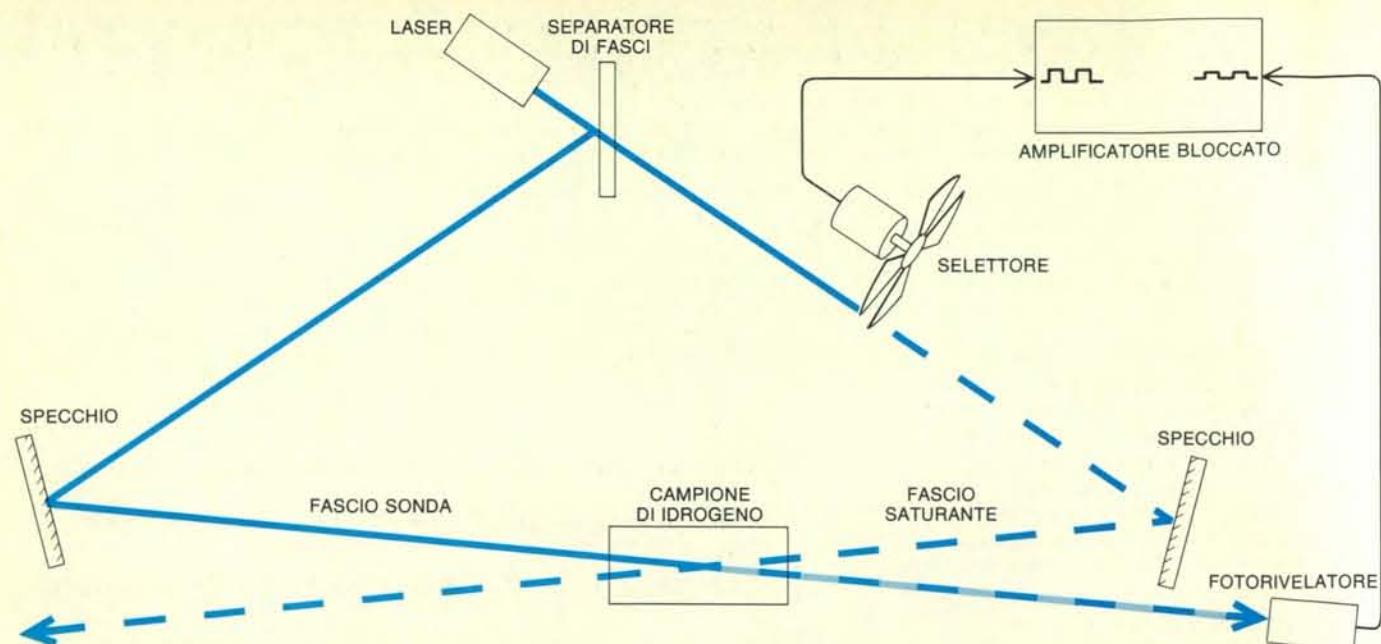
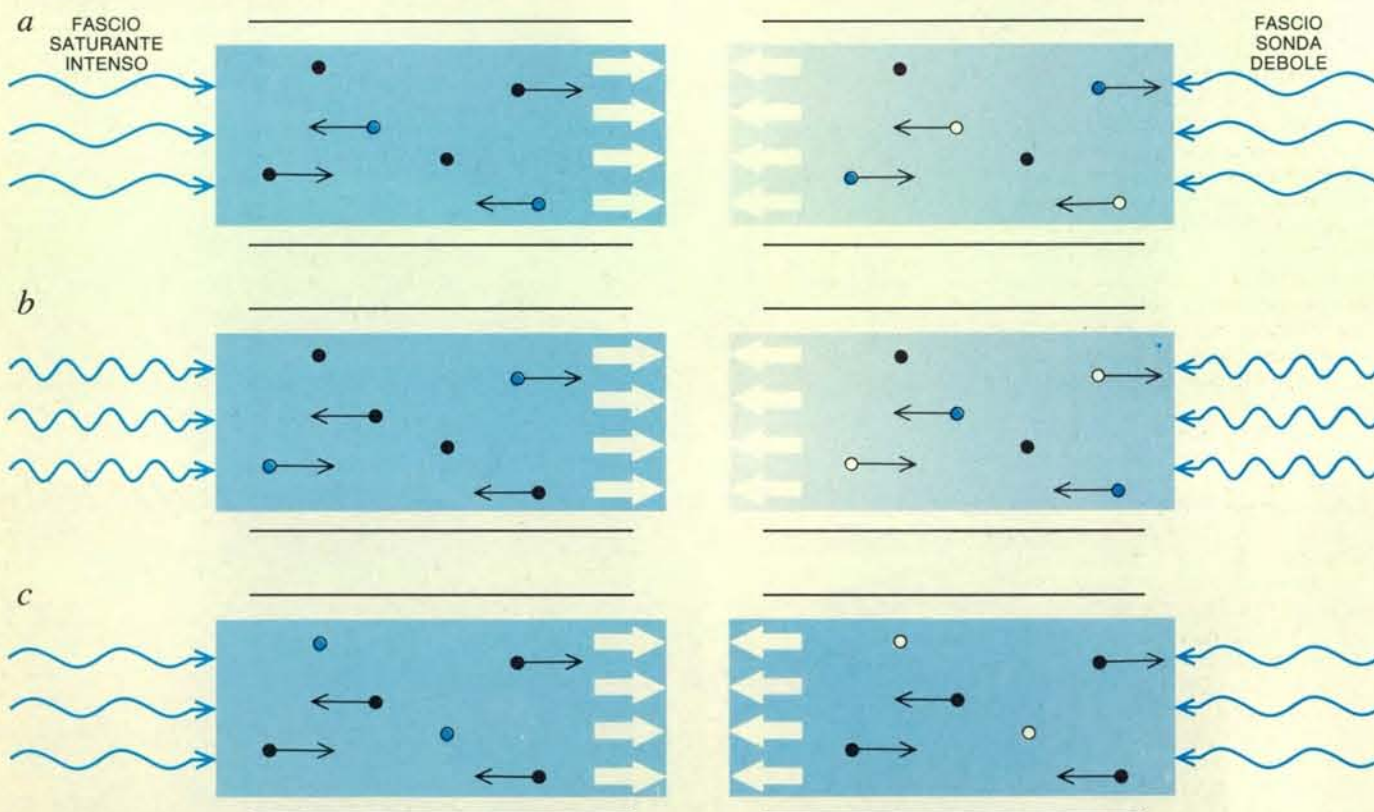
La questione venne definita nel 1947 in un brillante esperimento condotto da Willis E. Lamb, Jr. e Robert C. Retherford. Essi scoprirono che il livello $2S_{1/2}$ era spostato verso l'alto rispetto al livello $2P_{1/2}$ di circa 1060 megahertz, o milioni di cicli al secondo. Confrontata con la frequenza della riga Balmer-alfa, che è di circa 450 000 gigahertz (miliardi di cicli al secondo), è uno spostamento nella posizione della componente $2S_{1/2}$ di un fattore di poco più di due parti per milione.

Lamb e Retherford non tentarono di risolvere le due componenti nello spettro ottico, ma usarono invece energia a radiofrequenza per stimolare una transizione diretta tra i livelli $2P_{1/2}$ e $2S_{1/2}$. La frequenza delle radioonde è circa 500 000 volte minore della frequenza della riga Balmer-alfa; dato che l'allargamento Doppler della riga è ridotto dello stesso fattore, può venire trascurato. Questa piccola scissione dei livelli energetici è oggi chiamata *spostamento Lamb.*

La necessità di dare una spiegazione allo spostamento Lamb suggerì un'altra revisione fondamentale della teoria fisica, introdotta da Richard P. Feynman, Julian Schwinger e Sin-itiro Tomonaga. Il loro punto di partenza era la nuova idea che gli elettroni osservati negli esperimenti possono differire in alcune proprietà, quali la massa e la carica, dall'ipotetico elettrone «nudo» che si pensa esistere in un universo vuoto. Questa differenza, precedentemente ignorata, può modificare le proprietà di un elettrone in un sistema legato. La teoria basata su queste idee è chiamata elettrodinamica quantistica.

In elettrodinamica quantistica lo spostamento Lamb è il risultato di un'interazione tra l'elettrone e le «fluttuazioni di punto zero» del campo elettromagnetico. Tali fluttuazioni, che sono sempre presenti, fanno «vagabondare» l'elettrone in modo da fargli apparire il nucleo come un qualcosa disperso nello spazio. Il risultato è un'apparente alterazione della forza tra le due particelle quando sono vicine. Dato che l'elettrone passa più tempo vicino al nucleo nello stato S che non nello stato P, l'energia nello stato S è leggermente maggiore.

Questo modello dell'atomo di idrogeno presenta un'ulteriore complicazione. Se L'elettrone è considerato una particella puntiforme, esso può essere soggetto a fluttuazioni di frequenza infinitamente alta, caratterizzate da energia infinita. Dal momento che gli infiniti vengono normalmente evitati, nasce la necessità di rispondere a domande sulle dimensioni e sulla struttura dell'elettrone, confrontando un elettrone legato in un atomo con un elettrone ipoteticamente nudo. La piccola energia misurata viene calcolata per sottrazione. Tali calcoli sono complessi, ma le previsioni della teoria sono eccezionalmente precise. In effetti l'elettrodinamica quantistica è la teoria più precisa mai proposta e non è stata trovata alcuna discrepanza con gli esperimenti.

La scissione di struttura fine di vari li-

Il laser a colorante emette luce confinata in un intervallo ristretto di lunghezze d'onda, ma può essere accordato anche in un ampio intervallo. Il mezzo laser è un colorante fluorescente che emette luce con uno spettro largo e praticamente continuo quando il colorante viene «pompato» da un altro laser. Da questa larga banda di emissione viene selezionato un intervallo di lunghezze d'onda molto più ristretto per essere amplificato in una cavità ottica accordata per risuonare a una frequenza particolare. Una accordatura grossolana viene fornita da un reticolo di diffrazione, che disperde la luce in modo da riflettere nella cella a colorante solo le lunghezze d'onda selezionate. Un interferometro campione restringe ulteriormente la larghezza di banda lasciando passare solamente quelle lunghezze d'onda che interferiscono costruttivamente in molte riflessioni interne. Un telescopio posto nella cavità allarga il fascio in modo da illuminare molte incisioni del reticolo. La luce monocromatica ma accordabile è stata impiegata in misure di spettroscopia per ottenere l'eliminazione dell'allargamento Doppler.

La spettroscopia in saturazione elimina l'allargamento Doppler selezionando un gruppo di atomi che non hanno alcuna componente del moto lungo l'asse ottico. La luce proveniente da un laser a colorante viene separata in un intenso fascio saturante e in un fascio d'onda più debole, che attraversano un campione di idrogeno atomico in versi opposti. Il fascio saturante è abbastanza intenso da ridurre in modo significativo la popolazione di atomi nello stato capace di assorbire la lunghezza d'onda del laser: il fascio lascia nel gas un solco privo di atomi. Il fascio sonda incontra perciò un minore assorbimento e fa registrare sul rivelatore una intensità più alta. I due fasci possono però interagire in questo modo solo se vengono assorbiti dagli stessi atomi del gas e ciò può accadere se vengono accordati sulla lunghezza d'onda di atomi che non subiscono spostamenti Doppler. In pratica il fascio saturante è interrotto da un «selettore» meccanico e un aumento di intensità del fascio sonda è rivelato accordando il laser su un intervallo di lunghezze d'onda e cercando un segnale alla frequenza del selettore.

I fasci che si propagano in versi opposti emessi da un laser a colorante accordabile subiscono spostamenti Doppler opposti. Quando la frequenza del fascio saturante (misurata nel sistema di riferimento del laboratorio) è inferiore a quella di alcune componenti selezionate nello spettro *(a)*, solo gli atomi in moto verso sinistra possono assorbire la luce, perché il loro movimento fa spostare la frequenza fino a portarla in risonanza con la transizione atomica. Il fascio sonda, con la stessa frequenza, viene assorbito soltanto dagli atomi che si muovono verso destra e non subisce alcuna influenza da parte del fascio saturante. Quando la frequenza del laser è maggiore di quella della componente spettrale *(b)*, il fascio saturante viene assorbito dagli atomi che si muovono verso destra e il fascio sonda dagli atomi in moto verso sinistra, che costituiscono ancora gruppi diversi di atomi. Soltanto quando la frequenza del laser coincide esattamente con la frequenza della transizione atomica *(c)* i due fasci interagiscono con gli stessi atomi: quegli atomi che sono effettivamente in quiete. Il fascio sonda scopre allora che gli atomi che avrebbero potuto assorbirlo non lo possono fare perché hanno già assorbito radiazioni dal fascio saturante. L'intensità del fascio sonda trasmesso pertanto aumenta, mentre scompare solamente la componente del moto diretta lungo l'asse ottico.

velli energetici dell'atomo di idrogeno è stata rivelata nei minimi particolari con misure a radiofrequenza. Contemporaneamente il miglioramento delle tecniche spettroscopiche ha condotto a determinazioni esatte della costante di Rydberg mediante misure della lunghezza d'onda di una componente della riga Balmer-alfa. Bryan P. Kibble, William R. C. Rowley, Raymond E. Shawyer e uno di noi (Series) hanno impiegato una sorgente luminosa raffreddata a elio liquido per determinare la costante di Rydberg con una precisione superiore a una parte su 10 milioni. Anche lo spostamento Lamb è stato osservato da uno di noi nel 1948 (Series) e analoghi spostamenti sono stati misurati in elio ionizzato da uno di noi (Series), da Gerhard Herzberg e da Frederick L. Roesler. Tuttavia la precisione degli esperimenti ottici non raggiunge quella dei metodi a radiofrequenza.

Larghezza delle righe spettrali

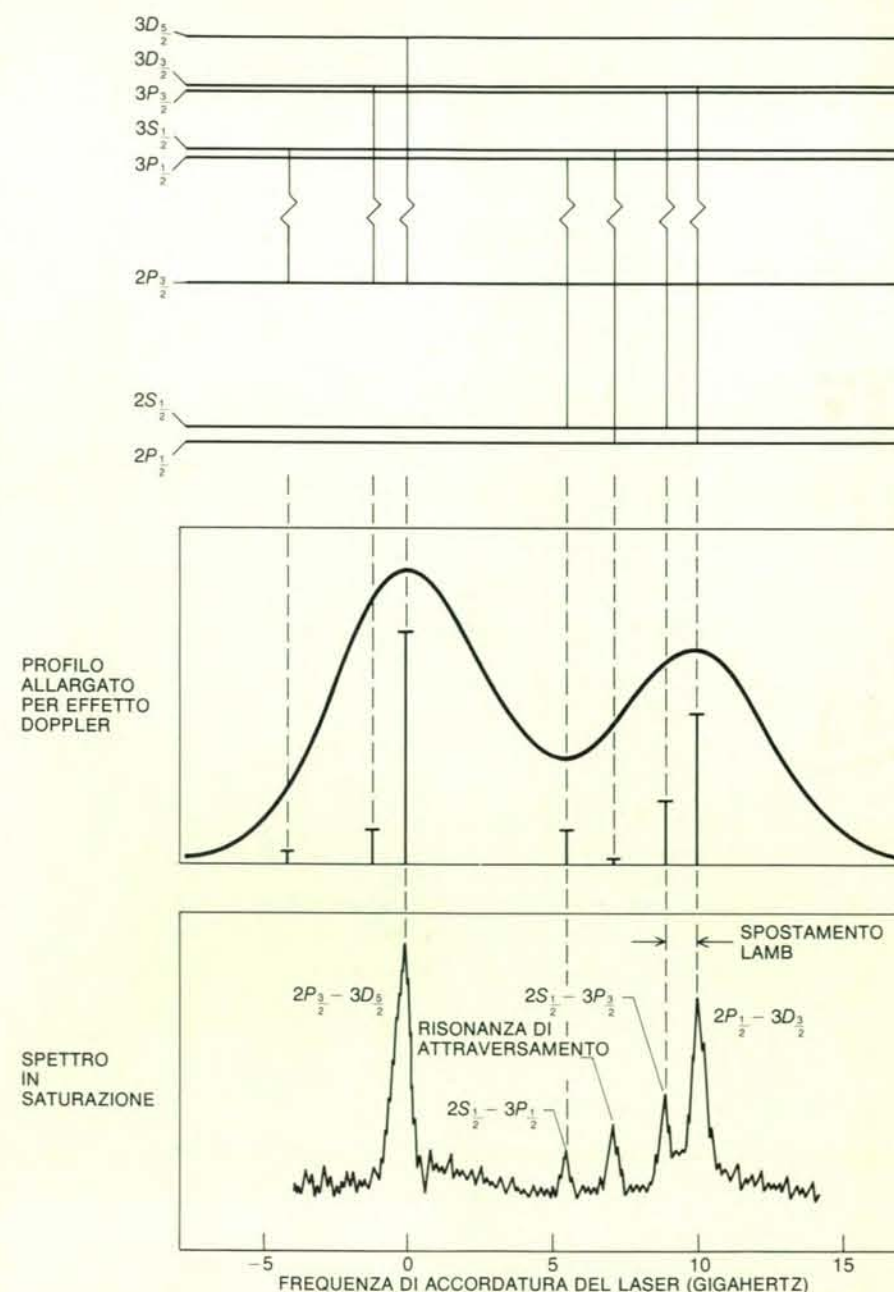
Benché questi esperimenti abbiano avuto un notevole successo, l'atomo di idrogeno costituisce una tale base di verifica fondamentale per le leggi della fisica che è importante analizzarlo più profondamente man mano che si hanno a disposizione nuove tecniche. Per ora il metodo più promettente è quello di migliorare la risoluzione degli spettri misurati; è la risoluzione che pone un limite alla nitidezza delle righe spettrali osservate e quindi alla precisione con cui si possono determinare le loro lunghezze d'onda.

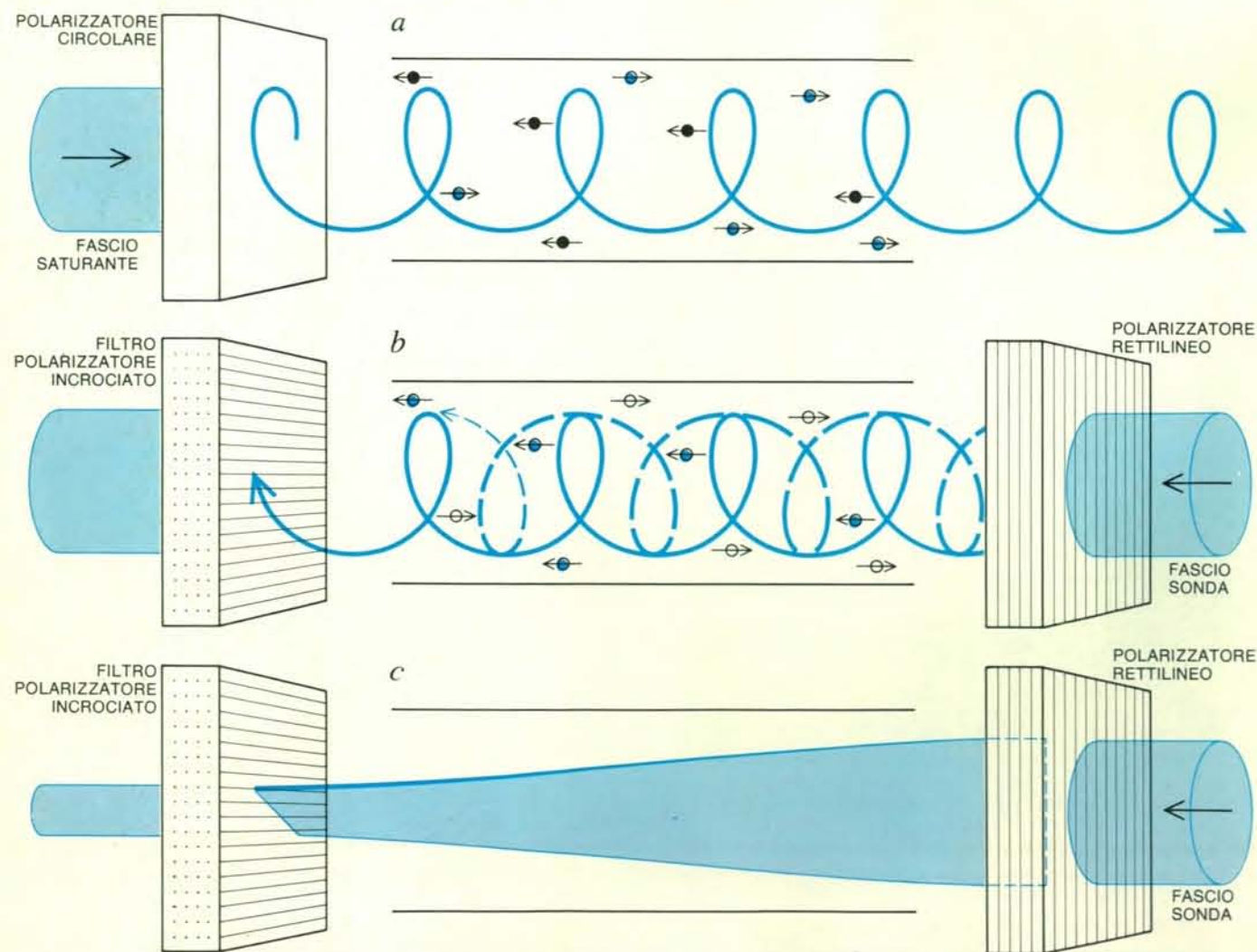
I primi studi spettroscopici degli atomi erano limitati dalla risoluzione degli stessi strumenti, che erano un prisma o un reticolo di diffrazione, cioè una lastra di vetro recante molte incisioni vicinissime. Il reticolo ha l'effetto di diperdere la luce riflettendo le diverse lunghezze d'onda a diversi angoli. Per esempio, se l'intero spettro visibile venisse disperso in 15 gradi, alcune componenti della riga Balmer-alfa si potrebbero separare per meno di 0,05 secondi di arco.

Con l'applicazione degli interferometri all'analisi delle righe spettrali, venne eliminata la limitazione intrinseca dovuta agli strumenti. L'interferometro inventato da Michelson e un altro tipo progettato dagli ottici francesi Charles Fabry e Alfred Perot sono strumenti sensibili per il confronto di due lunghezze d'onda. Essi sono capaci di distinguere due picchi netti con frequenze differenti di un gigahertz o meno ancora. Gli interferometri non riescono a separare completamente le diverse componenti delle righe Balmer non perché siano inadeguati, ma perché le componenti stesse non sono ben separate. Ogni componente è distribuita su un intervallo di lunghezze d'onda spesso maggiore della distanza tra le componenti.

Una parte dell'allargamento è la larghezza «naturale» della riga spettrale. Perfino nelle migliori condizioni di osservazione nessuna componente dello spettro può essere assolutamente monocromatica. Tale limitazione deriva dal fatto che gli «stati stazionari» di un atomo non sono veramente stazionari. Infatti gli atomi irraggiano energia dopo un intervallo di tempo finito, e cioè gli stati eccitati finiscono col decadere.

La larghezza naturale di riga pone un limite ultimo alla risoluzione spettroscopica, ma fino a poco tempo fa tale limite è stato raramente raggiunto. Ciò perché le righe si allargano molto di più per un altro effetto: lo spostamento di frequenze Doppler dovuto al moto degli atomi.

Un metodo per diminuire notevolmente l'allargamento Doppler sta nel misurare non lo spettro di un gas, ma di un fascio atomico collimato. Un fascio del genere si può produrre lasciando sfuggire gli atomi di un gas da un recipiente attraverso una serie di piccolissimi fori. Tutti gli atomi che passano dai forellini si muovono nella stessa direzione e in tal modo la luce che attraversa il fascio perpendicolarmente subirà soltanto modesti spostamenti Doppler. Tuttavia il numero di atomi del fascio diminuisce al migliorare della collimazione. Inoltre, non è sempre possibile preparare un fascio di atomi nello stato necessario. Per esempio, per studiare lo spettro Balmer dell'idrogeno in assorbi-

La riga Balmer-alfa viene risolta in una serie di picchi netti con la spettroscopia in saturazione. I calcoli teorici indicano che la riga ha sette componenti risultanti dalle scissioni a struttura fine e dallo spostamento Lamb, ma trascurando la struttura iperfine di entità molto minore. In un profilo della riga con allargamento Doppler si possono distinguere solo due componenti. Esaminando soltanto quegli atomi che non hanno alcuno spostamento Doppler, la spettroscopia in saturazione consente di risolvere chiaramente quattro componenti. (Un quinto picco che è posto a metà strada tra altri due picchi ha origini spurie e viene chiamato usualmente risonanza di attraversamento.) La scala della frequenza misura l'accordatura del laser da un punto di partenza arbitrario.

La spettroscopia in polarizzazione si basa sul fatto che si possono più facilmente rivelare piccole variazioni nello stato di polarizzazione della luce che piccole variazioni di intensità. Il fascio uscente da un laser anche in questo caso viene separato in due fasci, ma il fascio saturante viene polarizzato circolarmente, mentre il fascio sonda presenta una polarizzazione rettilinea. Dato che il fascio sonda incontra filtri polarizzatori incrociati, soltanto una parte di esso può raggiungere il rivelatore, a meno che l'idrogeno gassoso possa in qualche modo far ruotare il piano di polarizzazione del fascio. Una tale rotazione può essere provocata dal fascio saturante polarizzato circolarmente, ma solo se i due fasci interagiscono con la stessa popolazione di atomi. In questo modo gli atomi che in un dato istante si trovano in quiete sono selezionati, come già nella spettroscopia in saturazione, in base alla loro capacità di interagire con onde luminose che si muovono in direzioni opposte.

La variazione dello stato di polarizzazione di un fascio sonda viene indotta dal precedente passaggio di un fascio saturante. Dato che il fascio saturante è polarizzato circolarmente, esso viene assorbito in modo selettivo dagli atomi che hanno una particolare orientazione *(a)*, e tali atomi vengono perciò eliminati dalla popolazione di atomi che possono assorbire radiazioni alla lunghezza d'onda del laser. La maggior parte degli atomi residui ha orientazione opposta. Il fascio sonda polarizzato rettilineamente può essere considerato come una combinazione di due onde polarizzate circolarmente in verso opposto, ma con la stessa intensità *(b)*. Quando il fascio sonda attraversa il campione, una delle componenti polarizzate circolarmente viene assorbita più intensamente, avendo gli atomi una orientazione preferenziale. L'assorbimento selettivo di una componente polarizzata circolarmente determina una variazione dello stato di polarizzazione della luce residua *(c)* e quindi una parte del fascio sonda può penetrare nel filtro analizzatore. La variazione dello stato di polarizzazione può verificarsi, però, solamente quando i due fasci vengono assorbiti dalla stessa popolazione di atomi, cioè da quegli atomi che non subiscono alcuno spostamento Doppler.

mento, gli atomi si devono trovare nello stato $n=2$ e tali atomi non si possono produrre in fasci abbastanza densi per le misure spettroscopiche convenzionali.

Le tecniche di spettroscopia laser introdotte fin dal 1970 hanno sensibilmente migliorato la risoluzione dal momento che comprendono metodi per l'eliminazione dell'allargamento Doppler. Discuteremo qui tre di tali metodi. Due di essi si basano sulla scelta dell'analisi di quei soli atomi di un gas la cui componente del moto lungo l'asse ottico risulta nulla. Il terzo metodo tiene conto del contributo di tutti gli atomi di un gas, ma elimina l'allargamento delle velocità mediante due spostamenti Doppler in versi opposti.

Il laser a colorante accordabile

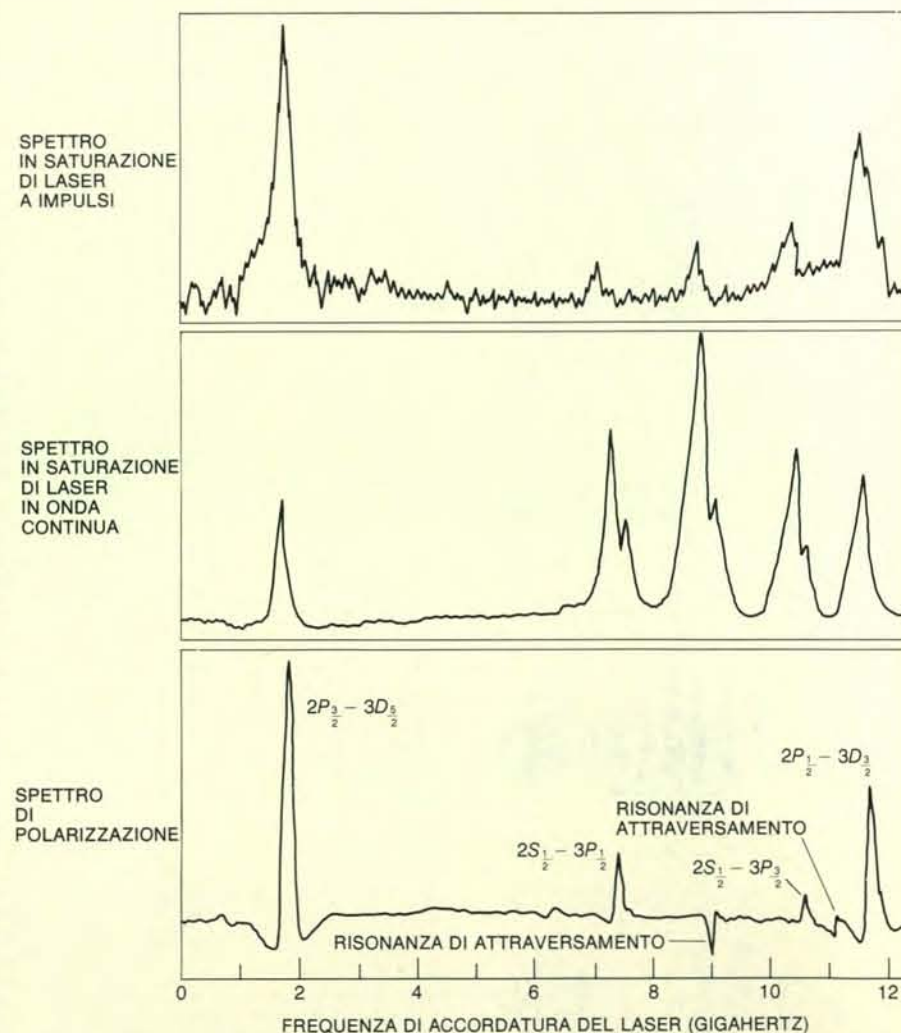
I laser forniscono luce di luminosità, direzionalità e purezza spettrale inugua-gliabili. Tuttavia i primi laser potevano emettere luce soltanto ad alcune lunghezze d'onda discrete determinate dalle righe spettrali degli atomi o ioni attivi. Uno strumento siffatto poteva essere utile in spettroscopia soltanto se la transizione in esame risultava coincidere con una delle righe laser disponibili.

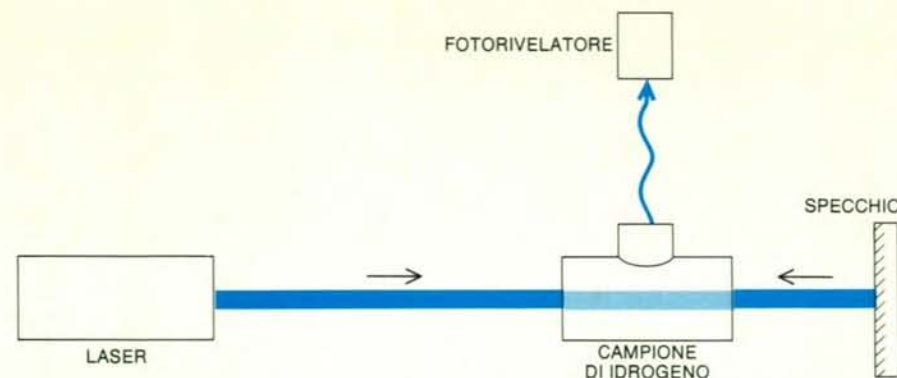
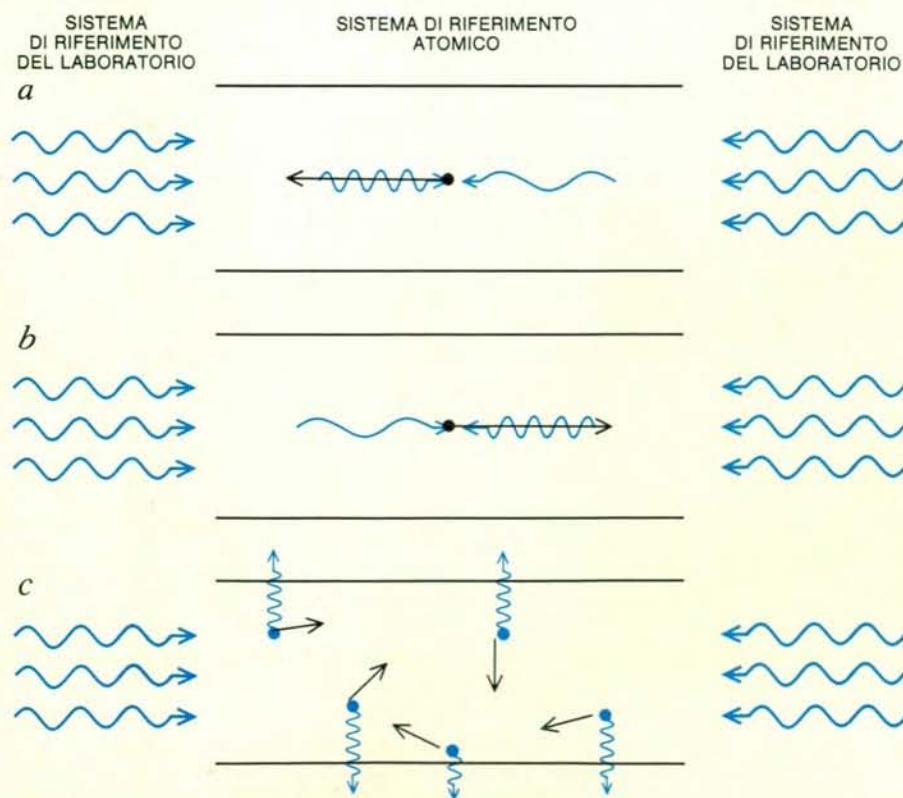
Nel 1966 Peter Sorokin del Centro di ricerca Thomas J. Watson dell'International Business Machines Corporation e, indipendentemente, Fritz Schäfer dell'Università di Marburgo scoprirono che soluzioni liquide di certe molecole organiche fluorescenti possono servire come mezzo attivo per un laser con una banda di lunghezze d'onda larga e continua. Le molecole organiche sono coloranti, dotati per loro natura di intense bande di assorbimento; è questo il motivo per cui sono fortemente colorate. Nel laser le molecole del colorante sono «pompate» o portate in uno stato eccitato da un'intensa sorgente esterna di luce, quale un altro laser.

Benché il laser a colorante sia in grado di amplificare la luce in un ampio intervallo di lunghezze d'onda, si scoprì ben presto che si poteva scegliere una particolare lunghezza d'onda progettando un risonatore ottico che consentisse soltanto alla lunghezza d'onda prescelta di passare ripetutamente attraverso la cella amplificatrice a colorante. Nel 1970 uno di noi (Hänsch) progettò un semplice laser a colorante a impulsi con elevata purezza spettrale. Il colorante, rinchiuso in una cella di vetro, veniva pompato con impulsi di radiazione ultravioletta provenienti da un laser ad azoto. Il risonatore era formato da una lastra di vetro parzialmente riflettente posta a un'estremità del laser e da un reticolo di diffrazione all'altra. Il reticolo disperde la luce emessa dal colorante con il risultato che viene riflesso all'indietro lungo l'asse ottico soltanto un ristretto intervallo di lunghezze d'onda. Un telescopio posto nella cavità laser allarga il fascio prima che questo incida sul reticolo, in modo da illuminare con luce ben collimata molte ondulazioni. La larghezza di riga del laser viene ulteriormente ridotta dall'inserimento di un interferometro Fabry-Perot inclinato nella cavità. L'interferometro è costituito da due superfici parzialmente riflettenti la cui spaziatura è controllata con esattezza; il tipo di interferometro inserito nel risonatore laser è chiamato *etalon* ed è costituito da una sola lastra di vetro le cui due superfici parallele sono rivestite con strati parzialmente riflettenti. Le onde luminose con lunghezze d'onda differenti seguono nel campione percorsi leggermente differenti. In conseguenza di riflessioni interne multiple alcune lunghezze d'onda interferiscono distruttivamente e si annullano, mentre altre interferiscono costruttivamente e si rinforzano.

La larghezza di riga di questo laser è di poche centinaia di nanometri, ovvero circa una parte per milione, confrontabile con la risoluzione dei più grandi spettrografi a reticolo. La larghezza può essere ulteriormente ridotta di circa un fattore 100 con un secondo interferometro Fabry-Perot disposto all'esterno della cavità laser, che si comporta come un filtro passivo. Anche se in ogni disposizione la luce è altamente monocromatica, il laser può essere accordato su tutto l'intervallo del profilo di assorbimento del colorante. Una regolazione grossolana si può ottenere variando l'angolo del reticolo di diffrazione o del campione, mentre si possono eseguire regolazioni più fini rinchiudendo il reticolo e il campione in una camera a tenuta d'aria e modificando la pressione con il risultato di cambiare l'indice di rifrazione del gas contenuto.

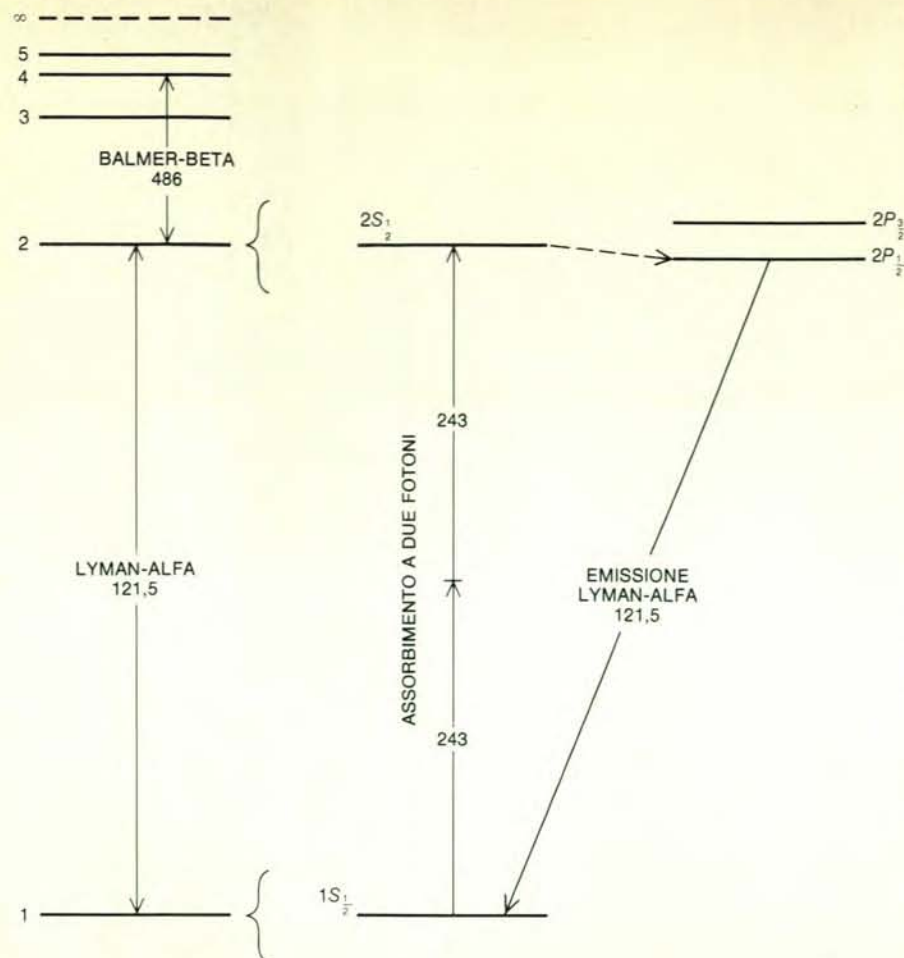
Spettroscopia in saturazione

Esiste un limite invalicabile alla risoluzione in lunghezze d'onda di qualsiasi laser a impulsi. La precisione con cui si può conoscere la lunghezza d'onda è determinata dal numero di volte in cui la misura

Uno spettro della riga Balmer-alfa è stato registrato con più elevata risoluzione con la spettroscopia in polarizzazione. Per confronto viene riportato in alto lo spettro in saturazione mostrato a pag. 103, mentre al centro è riportato uno spettro in saturazione migliorato, ricavato con un laser in onda continua anziché con un laser a impulsi. Nello spettro in polarizzazione non si possono rivelare componenti addizionali, ma i quattro picchi risolti sono molto più netti. I due picchi piccoli posti tra quelli più alti (escludendo le risonanze di attraversamento) hanno una larghezza misurata di circa 35 megahertz, circa 10 volte più stretta delle righe corrispondenti nello spettro in saturazione a impulsi. Lo spettro in polarizzazione ha un aspetto insolito (oscilla sotto e sopra il livello fondamentale) perché il grafico presenta non l'intensità della luce trasmessa, ma il rapporto di variazione di intensità al variare della regolazione del laser lungo il suo intervallo di frequenze.

può essere ripetuta e quindi dalla durata dell'impulso. Righe più strette richiedono un laser in onda continua. Nel 1970 Otis G. Peterson, Sam A. Tuccio e Benjamin B. Snavely dei Laboratori di ricerca della Eastman Kodak misero in funzione il primo laser a colorante accordabile funzionante in onda continua. I suoi principi di funzionamento erano gli stessi del laser a colorante a impulsi, ma la soluzione del colorante veniva fatta scorrere rapidamente per evitarne il surriscaldamento. Oggi esistono in commercio laser a colorante in onda continua altamente perfezionati. In molti laboratori si sono raggiunte larghezze di riga fino a poche parti su 10^{12} con stabilizzazione elettronica della frequenza laser.

La spettroscopia a due fotoni non seleziona atomi non soggetti a spostamenti Doppler, ma annulla gli spostamenti Doppler di tutti gli atomi di un gas. Il laser a colorante viene accordato su una frequenza che è esattamente la metà di quella di una particolare transizione atomica. Il fascio attraversa il campione e viene riflesso all'indietro, determinando la formazione di un campo di onde stazionarie. Gli atomi del gas possono subire la transizione prescelta assorbendo contemporaneamente due fotoni, o quanti di luce, provenienti da direzioni opposte e che perciò subiscono spostamenti Doppler opposti. Gli stati eccitati creati in questo modo decadono a un livello energetico inferiore, emettendo fotoni che possono quindi venire rivelati e contati.

Nella spettroscopia a due fotoni si annullano spostamenti Doppler opposti. Un atomo in moto verso sinistra *(a)* vede l'onda proveniente da sinistra spostata verso una frequenza maggiore, mentre la frequenza dell'onda proveniente da destra viene diminuita della stessa entità. Per gli atomi in moto verso destra *(b)* gli spostamenti di frequenza risultano opposti ai precedenti. Un atomo in quiete non subisce alcuno spostamento Doppler. Indipendentemente da quale sia la velocità dell'atomo, la somma delle due frequenze rimane costante e uguale al doppio della frequenza del laser (misurata nel sistema di riferimento del laboratorio). Tutti gli atomi possono quindi assorbire due fotoni *(c)* e riemetterli in seguito, a qualche frequenza più elevata, quando decadono.

Data una sorgente di luce monocromatica ma accordabile, si può misurare uno spettro di assorbimento facendo passare la luce attraverso un campione del gas e esaminando una per una le frequenze vicine a una riga spettrale. L'intensità della luce trasmessa può essere controllata da un rivelatore fotoelettrico, che dovrebbe rivelare un minimo in corrispondenza della lunghezza d'onda di ogni componente della riga. Questo schema, però, dovrebbe misurare semplicemente il profilo della riga allargato per effetto Doppler. Per poter utilizzare pienamente i vantaggi della miglior risoluzione del laser si deve trovare qualche metodo per eliminare gli spostamenti Doppler.

A uno di tali metodi è stato imposto il nome di spettroscopia in saturazione. Esso è basato su due proprietà della luce laser: piccola larghezza di riga e alta intensità. Quando un atomo assorbe un quanto di luce, esso viene eccitato in un livello energetico più elevato e viene pertanto temporaneamente escluso dall'insieme degli atomi che assorbono luce. In un fascio di luce intenso gli atomi di un gas possono assorbire quanti con maggior velocità di quella con cui possono ritornare al loro stato originale. Il risultato è che il numero degli atomi assorbitori è praticamente ridotto a zero. La luce attraversa il gas e un secondo fascio della stessa lunghezza d'onda trova un minor coefficiente di assorbimento del gas.

Lamb fu il primo a scoprire le possibili applicazioni di questo effetto ai fini della spettroscopia ad alta risoluzione. Egli fece notare che le due onde che viaggiano in versi opposti all'interno di un laser potrebbero cooperare per saturare l'assorbimento di una particolare categoria di atomi nel mezzo attivo del laser. Supponendo che la larghezza di riga della luce sia abbastanza piccola, essa deve sempre interagire soltanto con quegli atomi il cui moto sposta la lunghezza d'onda fino a portarla in risonanza con una componente dello spettro atomico. Se la frequenza della luce è minore di quella della riga componente, per esempio, entrambe le onde nella cavità interagiscono con gli atomi che vanno loro incontro e si tratterà di atomi differenti per i due fasci diretti in verso opposto. Esiste tuttavia un modo di accordare il laser per il quale entrambi i fasci interagiscono con gli stessi atomi: l'accordatura che corrisponde alla esatta lunghezza d'onda della componente spettrale, per la quale gli atomi che possono assorbire la luce sono quelli in quiete o quelli in moto trasversalmente all'asse ottico. Quando il laser viene accordato su questa lunghezza d'onda, la popolazione di atomi nello stato di assorbimento si riduce a zero e l'uscita del laser si annulla. Questo *Lamb dip* (minimo di Lamb) fu osservato per la prima volta da Ross A. McFarlane, William R. Bennett, Jr. e Lamb. Fu utilizzato per spettroscopia a elevata risoluzione da Abraham Szöke e Ali Javan del MIT, ma si è dimostrato di utilità limitata, poiché con esso si possono studiare solo le transizioni laser stesse o le righe molecolari che vengono a coincide-

Lo stato fondamentale, o di minima energia, dell'idrogeno si può raggiungere soltanto con transizioni le cui lunghezze d'onda appartengono alla regione dell'ultravioletto lontano dello spettro elettromagnetico; tali transizioni si possono osservare più facilmente con la spettroscopia a due fotoni che con qualsiasi altra tecnica. La riga Lyman-alfa, una transizione da $n=1$ a $n=2$, ha una lunghezza d'onda di 121,5 nanometri e per la spettroscopia a due fotoni è necessaria una radiazione di lunghezza d'onda doppia, pari a 243 nanometri. Anche quest'ultima lunghezza d'onda non si può produrre con il laser a colorante e si può produrre raddoppiando la frequenza (o dimezzando la lunghezza d'onda) di un laser accordato a 486 nanometri. La duplicazione di frequenza viene ottenuta con un cristallo che, sottoposto a un'illuminazione intensa, emette la seconda armonica della frequenza incidente. La lunghezza d'onda originale del laser (486 nanometri) viene a coincidere all'incirca con la lunghezza d'onda della riga Balmer-beta. L'assorbimento di due fotoni nell'idrogeno atomico stimola soltanto la transizione $1S_{1/2}$ - $2S_{1/2}$. L'atomo eccitato viene portato attraverso piccoli urti con altri atomi allo stato $2P_{1/2}$, dal quale ritorna allo stato fondamentale con emissione di un fotone dalla lunghezza d'onda di 121,5 nanometri.

re con le lunghezze d'onda del laser a gas e qualsiasi gas che si vuole esaminare deve essere posto entro il delicato risonatore di un laser in onda continua.

Nel 1970 Christian Bordé dell'Università di Parigi e indipendentemente uno di noi (Hänsch) misero a punto una tecnica per la misura di spettri in saturazione esternamente al risonatore laser. La luce di un laser accordabile viene separata in un fascio intenso saturante e in un fascio sonda più debole. I due fasci attraversano poi lo stesso percorso in una cella a gas, ma in versi opposti. Se il laser è lievemente scordato rispetto alla frequenza di una transizione atomica, i due fasci interagiscono con atomi differenti e il fascio saturante non ha alcun effetto sull'assorbimento del fascio sonda. Quando il laser viene accordato in modo da essere assorbito da atomi stazionari, però, il fascio saturante vuota la popolazione di atomi nello stato prescelto e il fascio sonda subisce un minore assorbimento. In pratica il fascio saturante seleziona tutti gli atomi aventi componente del moto nulla lungo l'asse ottico e ignora tutti gli altri.

In pratica la variazione di assorbimento è piccola e viene facilmente oscurata dal rumore. Per poter distinguere il segnale dal fondo il fascio saturante viene tagliato o interrotto a una frequenza nota, mentre il fascio sonda viene lasciato in funzione in modo continuo. Il segnale può poi essere rivelato esaminando l'ampiezza del fascio sonda per qualsiasi modulazione alla frequenza di interruzione.

Due di noi (Hänsch e Schawlow) insieme a Issa S. Shahin, allora studente diplomato della Stanford University, esaminarono con questo metodo la riga Balmer-alfa. Gli atomi di idrogeno venivano eccitati al livello di assorbimento $n=2$ in un tubo a scarica nei gas a bassa pressione. I due fasci di luce di un laser a colorante a impulsi venivano fatti passare attraverso una sezione del plasma di scarica lunga circa 15 centimetri, dove mostravano un assorbimento di circa il 50 per cento.

Delle sette componenti a struttura fine previste in teoria, se ne osservarono quattro, le due più intense delle quali risolte molto più nettamente che in qualsiasi spettro precedente. Ma la cosa più importante è che per la prima volta venne osservato nello spettro di assorbimento ottico lo spostamento Lamb che separa i livelli $2S_{1/2}$ e $2P_{1/2}$.

Lo spettro in saturazione comprendeva anche un quinto picco, ma era un picco spurio chiamato risonanza di attraversamento, che compare a metà strada tra una qualsiasi coppia di transizioni aventi in comune lo stesso livello superiore o inferiore. Quando il laser viene accordato sulla frequenza intermedia tra queste due componenti della riga, gli atomi in moto in una direzione possono assorbire il fascio saturante, mentre quelli in moto in direzione opposta possono assorbire il fascio sonda. Il risultato è che l'assorbimento viene saturato anche dai due tipi di atomi in movimento.

La strettezza dei picchi nello spettro di saturazione fece pensare che si potesse determinare la costante di Rydberg con maggiore precisione misurando la lunghezza d'onda assoluta di una componente della struttura fine. Una misura del genere venne eseguita da un gruppo di ricercatori di Stanford, sotto la guida di uno di noi (Hänsch), di cui faceva parte in particolare Munir Nayfeh. La componente scelta fu la transizione $2P_{3/2}$ - $3D_{5/2}$, che è la più intensa ed è quella con la minor scissione iperfine non risolta e con la minor perturbazione da parte del campo elettrico del tubo di scarica. In primo luogo fu necessario esaminare gli effetti, sulla posizione della riga, delle variazioni di pressione del gas, di corrente e di tensione nel tubo di scarica e di intensità del laser ed effettuare correzioni per tutti gli errori sistematici. Per poter misurare la posizione assoluta della riga anziché l'intervallo tra due righe fu necessario adottare una precisa lunghezza di riferimento, individuata in un laser elio-neo la cui frequenza venne fissata elettronicamente a una particolare componente iperfine di una riga di assorbimento di vapori molecolari di iodio; questa riga venne anche individuata con la spettroscopia in saturazione. La lunghezza d'onda del laser elio-neo stabilizzato era nota con precisione in funzione del campione internazionale di lunghezza, che è una riga spettrale del cripto emessa in particolari condizioni. I dati furono stimati nel 1974. Il risultato, $R = 10\,973\,731{,}43 \pm 0{,}10$ metri alla meno uno, fu più preciso del valore precedente di quasi un fattore 10.

Un metodo simile di spettroscopia esente da effetto Doppler venne introdotto nel 1976 da Carl Wieman, uno studen-

te diplomato di Stanford che lavorava con uno di noi (Hänsch). Basandosi sul fatto che piccole variazioni di polarizzazione della luce potevano essere rivelate più facilmente di variazioni di intensità, tale tecnica ha raggiunto una sensibilità decisamente migliore. Con una minore intensità del laser si possono osservare meno atomi, evitando quindi gli spostamenti e gli allargamenti di riga sistematici.

Spettroscopia in polarizzazione

Come accadeva per la spettroscopia in saturazione, la luce laser viene scissa in due fasci, uno più intenso dell'altro, che attaversano il campione in direzioni opposte. In questo caso, però, il fascio sonda più debole vede il campione di gas inserito tra due filtri polarizzatori incrociati, in modo tale che sul rivelatore giunge poca luce fintantoché il campione non fa variare lo stato di polarizzazione del fascio sonda. Il fascio saturante è responsabile di tali variazioni sul campione, dal momento che viene prima fatto passare attraverso una lamina quarto d'onda, un pezzo di materiale birifrangente tagliato e lavorato in modo da produrre una polarizzazione circolare della luce. La luce è detta polarizzata circolarmente quando il campo elettrico ruota (in senso orario o antiorario) anziché oscillare in un piano come fa nel caso della luce polarizzata rettilineamente.

La probabilità che un atomo assorba luce polarizzata circolarmente dipende dall'orientazione del momento angolare degli atomi rispetto al campo rotante. All'inizio l'orientazione degli atomi è casuale, ma il fascio saturante svuota il gas di quegli atomi orientati in modo da assorbire la luce polarizzata circolarmente in un senso, lasciando un eccesso di atomi con senso di polarizzazione opposto. Quando il fascio sonda, che è polarizzato rettilineamente, attraversa la stessa regione di gas, gli atomi orientati ne possono modificare la propagazione.

Un'onda polarizzata rettilineamente può essere considerata come la sovrapposizione di due onde polarizzate circolarmente di uguale intensità, una ruotante in senso orario, l'altra in senso antiorario. Quando questo fascio attraversa il gas, incontra atomi che tendono ad assorbire soltanto una delle componenti polarizzate circolarmente, attenuandola e lasciando più intensa l'altra componente. Il risultato è che quando il fascio sonda emerge dal gas esso risulta polarizzato ellitticamente e l'asse dell'ellisse è ruotato rispetto al piano della polarizzazione rettilinea originale. Il fascio ha in tal modo acquistato una componente che è in grado di attraversare il filtro polarizzatore incrociato. Tutto ciò può accadere, però, soltanto se fascio saturante e fascio sonda interagiscono con gli stessi atomi, cioè quelli che non subiscono spostamenti Doppler.

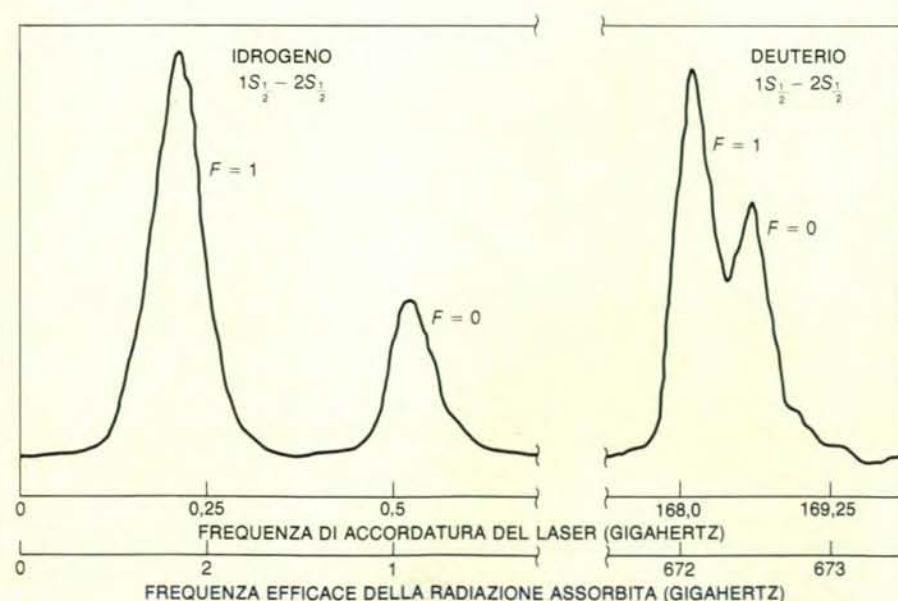
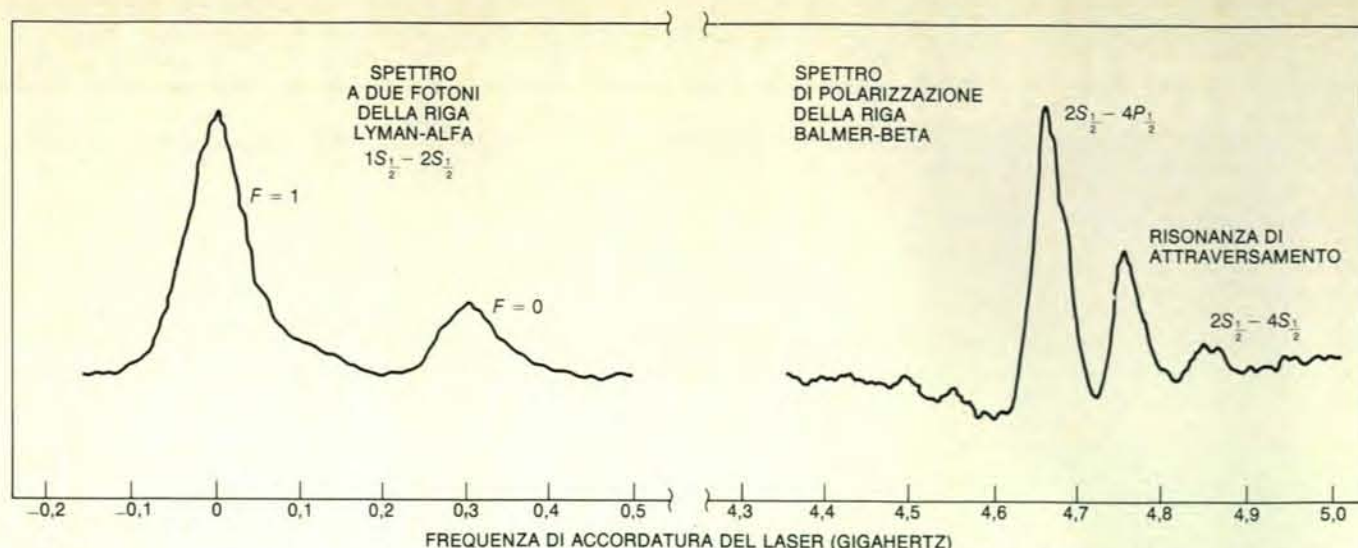
Dal filtro polarizzatore incrociato non passa luce se i due fasci interagiscono con differenti gruppi di atomi e quindi non c'è praticamente alcun segnale di fondo. Quindi i veri picchi dello spettro non vengono facilmente oscurati dal rumore o dalle fluttuazioni di intensità del laser. Quando il segnale è molto piccolo è spesso vantaggioso spostare lievemente i filtri polarizzatori dalla posizione incrociata. Il rivelatore registra in questo caso un segnale di fondo finito che può o aumentare o diminuire a seconda della direzione in cui ruota il piano di polarizzazione.

John Goldsmith, uno studente diplomato di Stanford, Erhard Weber, che si trovava a Stanford proveniente dall'Università di Heidelberg e uno di noi (Hänsch) hanno ottenuto recentemente una versione migliorata dello spettro della riga Balmer-alfa con la spettroscopia in polarizzazione laser. I fasci saturante e sonda erano prodotti da un laser a colorante a bassa potenza in onda continua con larghezza di banda minore di un megahertz.

Lo spettro veniva registrato sotto forma di derivata, una funzione matematica che misura la rapidità di variazione dell'intensità del fascio sonda quando il laser viene accordato su una banda di frequenze. Nello spettro di polarizzazione i due picchi più grandi sono più netti di quanto lo fossero nelle precedenti misure di spettroscopia in saturazione e si osservano altre tre componenti più piccole ancora più strette. La larghezza di riga di tali componenti è di circa 35 megahertz, quasi 10 volte più stretta che nelle misure precedenti. Queste tre componenti rappresentano tutte transizioni che partono dallo stato a lunga vita media $2S_{1/2}$ e appaiono tutte separate in due componenti iperfini distanti circa 178 megahertz. Sembrerebbe possibile risolvere chiaramente la scissione iperfine, ma si osservano invece soltanto picchi singoli. La spiegazione di questa discrepanza sta nel fatto che in tutti i casi uno degli stati iperfini ha un momento angolare totale nullo e non può pertanto essere polarizzato. Così, ognuno dei picchi osservati non è un miscuglio non risolto di due transizioni, ma una singola componente iperfine.

La componente della transizione $2S_{1/2}$ - $2P_{1/2}$ venne scelta per una nuova misura assoluta di lunghezza d'onda, calibrata ancora sulla riga di riferimento dello iodio assunta come campione di lunghezza. Il nuovo valore della costante di Rydberg ricavato attraverso questa misura è R = 10 973 731,476 ± 0,032 metri alla meno uno. La misura è quasi tre volte più precisa della precedente e pone la costante di Rydberg tra le costanti fisiche fondamentali note con maggiore precisione. Soltanto la velocità della luce e il rapporto tra i momenti magnetici dell'elettrone e del protone sono stati determinati con precisione confrontabile.

Spettroscopia a due fotoni

Questi metodi forniscono un conveniente approccio alle righe della serie di Balmer, quelle generate da transizioni tra il livello energetico $n = 2$ e i livelli superiori. Altri stati non sono raggiungibili con la stessa facilità. Uno particolarmente importante è lo stato $S_{1/2}$ con $n = 1$, che è lo stato fondamentale dell'atomo di idrogeno. Anche il più vicino gruppo di stati, quelli con $n = 2$, è talmente lontano dallo stato fondamentale che la lunghezza d'onda della riga corrispondente, chiamata Lyman-alfa, cade nella regione dell'estremo ultravioletto dello spettro elettromagnetico. A tali lunghezze d'onda, poste ben al di sopra dell'intervallo del laser a colorante, la spettroscopia di pre-

La struttura iperfine della transizione $1S_{1/2}$ - $2S_{1/2}$ è stata risolta con la spettroscopia a due fotoni sia nell'idrogeno normale che nel deuterio. Le transizioni tra i due livelli iperfini dell'idrogeno sono responsabili della radiazione di lunghezza d'onda di 21 centimetri osservata dai radioastronomi. La larghezza naturale di riga delle componenti è probabilmente molto piccola, forse dell'ordine di un hertz; qui le componenti vengono risolte a circa 100 megahertz. Lo spostamento tra l'idrogeno e il deuterio è dovuto principalmente alla differenza di massa dei nuclei.

Per determinare l'entità dello spostamento Lamb per lo stato $1S_{1/2}$ dell'idrogeno si è eseguita una misura simultanea della transizione $1S_{1/2}$ - $2S_{1/2}$ e della riga Balmer-beta. La riga Balmer-beta è stata misurata con spettroscopia in polarizzazione alla lunghezza d'onda fondamentale del laser, 486 nanometri. La frequenza raddoppiata all'uscita dallo stesso laser è stata simultaneamente utilizzata per registrare uno spettro a due fotoni della transizione $1S_{1/2}$ - $2S_{1/2}$ a una lunghezza d'onda reale di 121,5 nanometri. Se i livelli energetici di Bohr fossero esatti, le due transizioni dovrebbero apparire alla stessa frequenza di accordatura del laser, ma le scissioni dei livelli energetici introducono una piccola differenza. Dalla misura della differenza tra le frequenze di accordatura del laser per queste due righe, si è calcolata l'entità dello spostamento Lamb dello stato $1S_{1/2}$con una precisione 50 volte superiore a quella di altre misure. Il valore per l'idrogeno risulta di 8161 ± 29 megahertz.

cisione risulta particolarmente difficile.

Tuttavia con laser intensi le difficoltà della spettroscopia nell'estremo ultravioletto si possono evitare eccitando ogni atomo con due fotoni, che forniscono collettivamente la necessaria energia di eccitazione. La probabilità che un atomo assorba simultaneamente due fotoni è piuttosto piccola solitamente, ma cresce con il quadrato dell'intensità della luce. La spettroscopia a due fotoni è particolarmente interessante per l'idrogeno perché si può evitare l'allargamento Doppler con un metodo ingegnoso proposto per la prima volta nel 1970 da L. S. Vasilenko, V. P. Chebotayev e A. V. Shishaev dell'Istituto di termofisica di Novosibirsk in Unione Sovietica. Il metodo è stato sperimentato per la prima volta nel 1974 da tre gruppi di ricercatori che lavoravano indipendentemente. Invece di scindere il fascio laser in due componenti, il fascio viene riflesso all'indietro da uno specchio, in modo da far viaggiare sullo stesso asse le onde che si muovono verso destra e quelle che si muovono verso sinistra. La cella di gas viene posta nel campo a onde stazionarie risultante da questa sovrapposizione. Il laser viene poi accordato su una frequenza pari alla metà di quella di una data transizione atomica. Alla frequenza opportuna ogni atomo del gas può assorbire contemporaneamente due fotoni provenienti da direzioni opposte.

Supponiamo che un atomo in moto lungo l'asse ottico in un campo a onde stazionarie incontri due fotoni che si avvicinano da direzioni opposte. Uno dei fotoni subisce uno spostamento Doppler verso il blu, cioè verso le frequenze più alte, ma l'altro fotone è spostato verso il rosso della stessa quantità. Per un atomo in quiete entrambi i fotoni hanno la stessa energia. In tutti i casi l'energia totale o la frequenza dei due fotoni assorbiti è costante, indipendentemente dalla velocità dell'atomo. Quando la frequenza del laser viene leggermente spostata dalla metà della transizione, nessuno degli atomi può assorbire i due fotoni che si muovono in direzioni opposte. (Viene emesso, però, un piccolo segnale di fondo, dagli atomi in moto che assorbono due fotoni che si propagano nella stessa direzione.) Quando la lunghezza d'onda del laser coincide esattamente con la transizione atomica, si osserva un netto aumento del numero di atomi eccitati. Il segnale esente da effetto Doppler viene fortemente evidenziato perché tutti gli atomi del campione possono contribuire a esso e non solo quelli in quiete. Da questo punto di vista il metodo differisce fondamentalmente sia dal metodo di spettroscopia in saturazione che da quello di spettroscopia in polarizzazione, nei quali si sceglie una popolazione di atomi esenti da spostamenti Doppler. Nella spettroscopia a due fotoni esente da effetto Doppler si fa in modo che uno spostamento Doppler sia compensato dall'altro.

Una transizione a due fotoni di particolare interesse è quella dallo stato $1S_{1/2}$ allo stato $2S_{1/2}$. Essendo quest'ultimo uno stato a lunga vita media, ci si aspetta che la riga sia particolarmente stretta. Le valutazioni teoriche fanno pensare a una larghezza naturale di riga di circa un hertz, pari a meno di una parte su 10^{15}, il che renderebbe la transizione una delle risonanze più strette mai osservate in qualsiasi sistema fisico. Lo stato $1S_{1/2}$ è di per sé interessante perché il suo spostamento Lamb è il maggiore previsto dall'elettrodinamica quantistica. Un atomo al livello energetico $n=1$ non può avere momento angolare orbitale e quindi non esiste nessuno stato P, nelle vicinanze, dal quale poter determinare il valore dello spostamento. Lo spostamento è semplicemente quello del livello energetico $1S_{1/2}$ dalla posizione che avrebbe in assenza di effetti dell'elettrodinamica quantistica. L'entità dello spostamento si può determinare soltanto con una misura assoluta di qualche componente spettrale, come la transizione $1S_{1/2}$ - $2S_{1/2}$.

Un'eccitazione a due fotoni di questa transizione è stata osservata nel 1974 da Siu Au Lee, Wieman e altri a Stanford. La riga Lyman-alfa ha una lunghezza d'onda di 121,5 nanometri; un fotone con energia pari alla metà di questo valore ha una lunghezza d'onda di 243 nanometri, ancora nell'ultravioletto e al di sopra della portata di laser a colorante accordabili. I ricercatori di Stanford fecero uso perciò di un laser a colorante funzionante a 486 nanometri, cioè il doppio della lunghezza d'onda necessaria. La luce intensa di questa sorgente veniva indirizzata su un cristallo che fungeva da duplicatore di frequenza. A causa dell'alta intensità del campo elettromagnetico applicato il cristallo non soltanto rimandava luce alla frequenza originale, ma era anche costretto a emettere multipli superiori, tra i quali la seconda armonica, a 243 nanometri. A tale lunghezza d'onda compariva circa il 2 per cento dell'energia.

La radiazione ultravioletta con frequenza raddoppiata veniva riflessa da uno specchio per formare un'onda stazionaria all'interno di un tubo a scarica nei gas a bassa pressione. Gli atomi di idrogeno che avevano assorbito due fotoni ritornavano allo stato fondamentale emettendo un solo fotone nell'estremo ultravioletto a una lunghezza d'onda di 121,5 nanometri. Questi fotoni sfuggivano attraverso una finestra laterale della camera e venivano rivelati da un fotomoltiplicatore.

Con questo sistema sono state misurate le transizioni $1S_{1/2}$ - $2S_{1/2}$ sia dell'idrogeno che del deuterio. Entrambe le righe hanno una grande scissione iperfine, ben risolta negli spettri a due fotoni. Le transizioni tra i due stati iperfini dell'idrogeno sono responsabili della radiazione da 21 centimetri che costituisce un segnale importante in radioastronomia. C'è anche una grande scissione tra le righe dell'idrogeno e le righe del deuterio, dovuta principalmente alla differenza delle masse nucleari. Questo spostamento isotopico è stato determinato con una precisione 1000 volte superiore a quella di misure precedenti. Un ulteriore miglioramento di un fattore 10 potrebbe fornire un nuovo valore dell'importante rapporto tra la massa dell'elettrone e quella del protone.

L'interesse maggiore per lo stato $1S_{1/2}$ è la determinazione dello spostamento Lamb nello stato fondamentale. Negli anni cinquanta Herzberg, che si trovava allora all'Osservatorio Yerkes dell'Università di Chicago, fu in grado di rivelare lo spostamento, ma la precisione con cui ne misurò il valore era soltanto di circa il 14 per cento. La spettroscopia a due fotoni ha fornito l'opportunità di una determinazione molto più precisa.

Il modo più diretto per misurare lo spostamento Lamb dovrebbe essere quello di determinare la lunghezza d'onda assoluta della transizione $1S_{1/2}$ - $2S_{1/2}$, ma il valore dello spostamento Lamb calcolato da tale lunghezza d'onda dipende anche dal valore che si assume per la costante di Rydberg. Il gruppo di Stanford ha trovato un metodo migliore sfruttando una coincidenza peculiare ma non causale: la lunghezza d'onda fondamentale del laser a colorante, cioè 486 nanometri, è quasi coincidente con la riga Balmer-beta. Se la formula di Bohr fosse corretta, la corrispondenza dovrebbe essere esatta: l'intervallo da $n=1$ a $n=2$ (Lyman-alfa) dovrebbe essere pari al quadruplo dell'intervallo da $n=2$ a $n=4$ (Balmer-beta). In realtà i livelli sono scissi e lievemente spostati dalle correzioni relativistica e della elettrodinamica quantistica. Le correzioni per i livelli $n=2$ ed $n=4$ sono note con precisione, e quindi un confronto delle due transizioni può determinare lo spostamento Lamb dello stato fondamentale.

Tre di tali misure sono state eseguite dal gruppo di Stanford, la più recente e più precisa delle quali da parte di Wieman. La riga di riferimento Balmer-beta è stata osservata con spettroscopia in polarizzazione, mentre lo stesso laser misurava contemporaneamente la riga Lyman-alfa con spettroscopia a due fotoni. Lo spettro Balmer-beta non si limitò a risolvere le componenti della riga a struttura fine, ma rivelò anche la scissione di tali componenti nel debole campo elettrico assiale del tubo a scarica nei gas. In linea di principio lo spostamento Lamb dello stato fondamentale si potrebbe determinare misurando la differenza tra la riga Balmer e la riga corrispondente alla transizione $1S$-$2S$ e confrontando poi il risultato con le previsioni della teoria di Dirac. Applicando poi le correzioni note alle posizioni dei livelli $n=2$ ed $n=4$, qualsiasi differenza residua si potrebbe attribuire allo spostamento Lamb dello stato $1S$. In pratica non fu così facile, dato che si dovette tener conto dei possibili errori sistematici e delle imperfezioni dell'apparecchiatura. Il risultato finale ottenuto fu di 8161 ± 29 megahertz. I risultati sperimentali sono ancora lontani dalla precisione dei calcoli teorici, che danno un valore di 8149 ± 0,08 megahertz, ma rappresentano un significativo passo in avanti rispetto alla prima misura di Herzberg.

Migliore risoluzione

Si devono superare molti ostacoli formidabili per avvicinarci alla larghezza naturale di riga di un hertz della transizione $1S_{1/2}$ - $2S_{1/2}$. Non è sufficiente ridurre ulteriormente la larghezza di banda del laser, perché ci sono altre cause di allargamento della riga. Una di esse è uno spostamento Doppler del secondo ordine, conseguenza della teoria della relatività ristretta. Un atomo in moto appare oscillare più lentamente di uno in quiete, e quindi l'atomo in moto ha in pratica una frequenza di risonanza inferiore indipendentemente dalla sua direzione di moto. I fasci in moto in versi opposti della spettroscopia a due fotoni non possono annullare questo spostamento, che dovrebbe essere di circa 50 chilohertz per atomi di idrogeno a temperatura ambiente.

Un'altra causa di allargamento di riga è il tempo di transito finito di un atomo in moto nel fascio laser. Dato che un atomo in movimento resta esposto alla luce laser soltanto per poco tempo, anche la luce proveniente da un laser in onda continua viene ricevuta dall'atomo come un impulso, la cui breve durata limita la risoluzione potenziale dello spettro. In linea di principio i due effetti si potrebbero ridurre rallentando gli atomi o, in altri termini, raffreddando il gas. A questo proposito è incoraggiante il fatto che Daniel Kleppner del MIT ha recentemente dimostrato che l'idrogeno atomico può essere raffreddato alla temperatura dell'elio liquido (4,2 kelvin) senza condensare, anche se le molecole dell'idrogeno biatomico condensano a 20 kelvin. Due di noi (Schawlow ed Hänsch) hanno recentemente dimostrato che la stessa luce laser potrebbe raffreddare un gas fino a bassa temperatura. Ciò può apparire paradossale, dal momento che una luce intensa solitamente cede calore, ma dobbiamo ricordare che la luce laser, essendo coerente, rappresenta uno stato a bassa entropia, o a basso disordine. Uno stato a bassa entropia può essere usato per generare una bassa temperatura.

In molti laboratori si sta oggi tentando di applicare la tecnica della spettroscopia a due fotoni alla transizione $1S_{1/2}$ - $2S_{1/2}$ del positronio, un «atomo» idrogenoide formato da un elettrone e dalla sua antiparticella, il positone. Le proprietà di questo atomo esotico sono previste con grande precisione dall'elettrodinamica quantistica e non è necessario applicare alcuna correzione per la struttura del nucleo. Però l'esperimento è difficile, avendo il positronio una vita media di soli 140 nanosecondi.

L'atomo di idrogeno è stato al centro dell'attenzione per quasi un secolo, ma le informazioni contenute in questo semplice sistema di particelle sono ben lungi dall'essere esaurite. Recenti teorie unificate dei campi, quali quella di Steven Weinberg dell'Università di Harvard e di Abdus Salam dell'Imperial College of Science and Technology, prevedono per l'idrogeno, e per altri atomi, lievi effetti che potrebbero apparire come piccole variazioni nella polarizzazione della luce emessa o assorbita. Le variazioni sono così piccole che esiste perfino il dubbio che sia possibile rivelarle, tuttavia alcuni gruppi di ricercatori stanno oggi cercando di misurarle con sufficiente precisione per verificare le teorie. Con l'evolversi della tecnologia delle sorgenti di luce coerente, possiamo con certezza prevedere che l'esplorazione dello spettro dell'idrogeno continuerà per decenni. Finora tutte le recenti scoperte tendono a confermare le previsioni della teoria. È tuttavia il caso di sottolineare che molte scoperte rilevanti della fisica del secolo XX sono dovute alla scoperta di piccole discrepanze tra teoria ed esperimento in questo stesso spettro. L'atomo di idrogeno può riservare ancora sorprese per il futuro.

Le antiche catapulte

Testi greci e romani pervenuti fino a noi rivelano il notevole livello di abilità matematica e tecnica che si espresse nello sviluppo di questi antichi esempi di lanciamissili balistici

di Werner Soedel e Vernard Foley

Nel 399 a.C. Dionisio il Vecchio, tiranno della colonia greca di Siracusa, in Sicilia, preparò la sua città a una lunga guerra con Cartagine intraprendendo un programma di ricerca e di sviluppo. Utilizzando tecniche oggi familiari, come la formazione di grandi *équipe* di specialisti, la divisione del lavoro per frazionare i compiti in unità facilmente padroneggiabili e il ricorso a incentivi finanziari e psicologici, Dionisio mirò chiaramente sin dal principio alla produzione di nuove armi. Da questo programma uscirono quadriremi e forse quinqueremi, navi da battaglia con quattro o cinque ordini di rematori, che disponevano di una potenza da applicare al loro sperone maggiore di quella delle normali triremi. I tecnici di Dionisio progettarono anche le prime catapulte.

Queste antiche macchine lançiavano probabilmente frecce utilizzando archi poco più potenti di quelli che potevano essere tesi da un uomo. Meccanizzando la tensione dell'arco e lo scatto della corda, però, gli inventori delle catapulte resero possibile la costruzione di archi molto più potenti. Pare che queste macchine venissero costruite con vari materiali; l'arco era composto, in particolare, da una parte centrale in legno, alla quale erano incollati, frontalmente, uno strato di tensione di tendini animali e posteriormente uno strato di compressione in corno. Infine l'arco flessibile raggiunse i limiti intrinseci delle sue possibilità e fu soppiantato da catapulte fondate sul principio della torsione. In questa impostazione fasci compatti di fibre elastiche fortemente tese venivano sottoposti a un'ulteriore tensione per mezzo di un braccio d'arco rigido quando l'arma veniva caricata prima del lancio. Per le corde che componevano questi fasci potevano essere usati pelo di cavallo o anche capelli umani, ma per ottenere prestazioni superiori venivano preferiti tendini di animali.

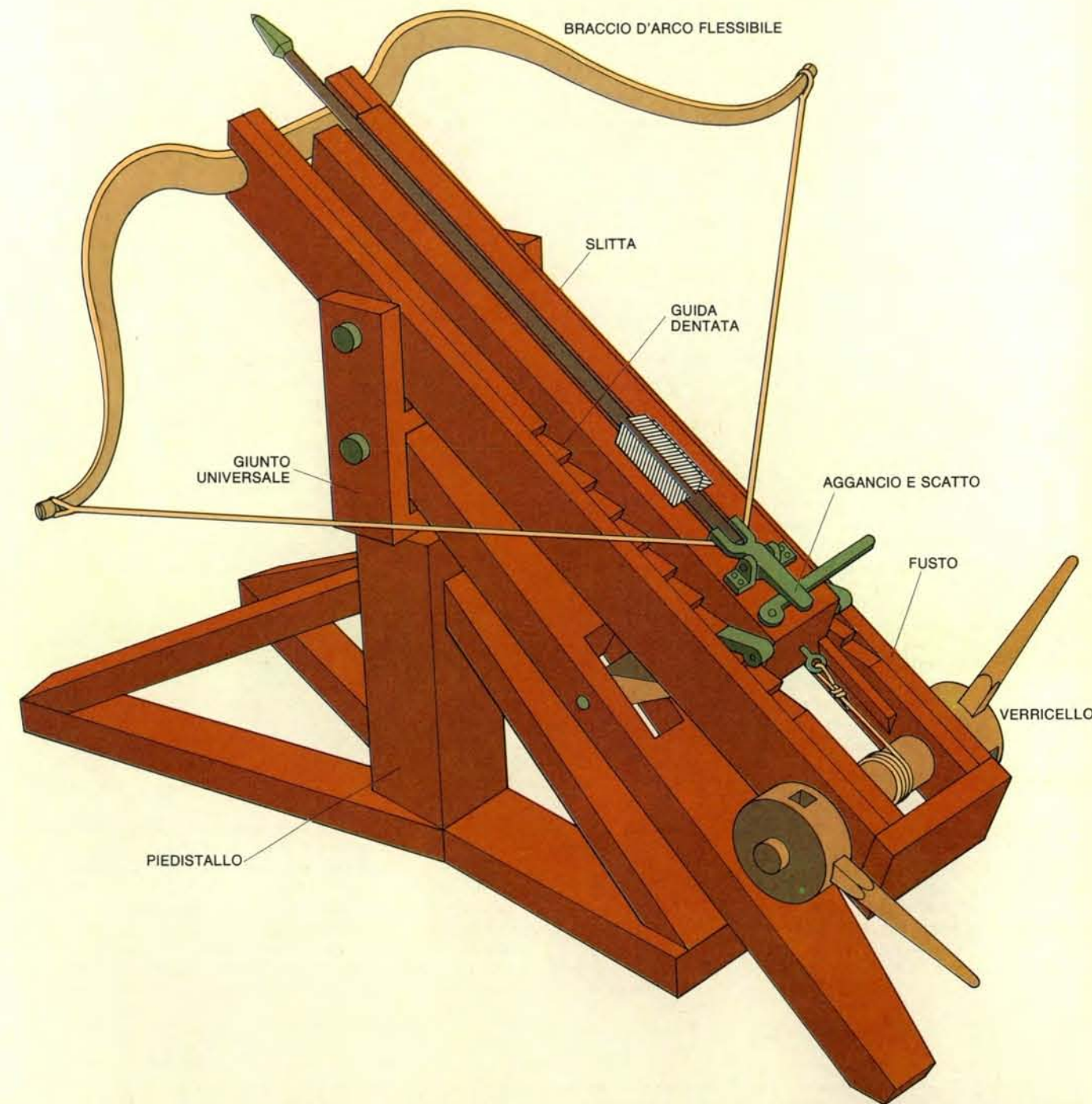
Per meccanizzare i movimenti della macchina, i progettisti delle catapulte incorporarono in esse un certo numero di importanti particolari costruttivi. L'elemento principale nella catapulta era il fusto, che formava l'asse principale dell'arma. Nella parte più alta di questo fusto composito c'era una guida a coda di rondine nella quale poteva scorrere avanti e indietro un altro fusto di dimensioni inferiori, la slitta. La slitta trasportava nella sua parte posteriore un dispositivo di aggancio e scatto, destinato appunto a bloccare e a liberare la corda dell'arco. Davanti a questo dispositivo c'era una scanalatura, sulla parte superiore della slitta, nella quale la freccia era alloggiata e dalla quale veniva lanciata. Per armare l'arco, si faceva avanzare la slitta finché il dispositivo di bloccaggio poteva agganciare la corda dell'arco. Poi la slitta veniva riportata all'indietro finché l'arco era teso completamente. Nelle prime versioni guide con denti di arresto disposte lungo il fusto impegnavano nottolini solidali alla slitta per resistere alla forza dell'arco. Più tardi fu adottata una ruota con denti di arresto nella parte posteriore del fusto. Nelle prime catapulte la slitta era probabilmente riportata indietro a mano, ma presto le dimensioni e la potenza di queste macchine richiesero l'uso di un verricello.

Non appena la catapulta divenne troppo grande per poter essere usata appoggiata a una spalla, venne collocata su un piedistallo. Allo scopo di facilitare il puntamento fu escogitato un giunto speciale per collegare il fusto col piedistallo. La soluzione di questo problema anticipò l'invenzione del giunto universale, la quale è attribuita di solito o a Girolamo Cardano (donde il nome di «giunto cardanico» o semplicemente «cardano») o a Robert Hooke, e viene quindi fatta risalire al Cinquecento o al Seicento. Analogamente, le superfici di scorrimento a coda di rondine della guida della slitta riappaiono ovunque nella costruzione di macchine moderne e l'uso di un gancio per sostituire l'azione della mano umana ha fondato una tradizione di manipolazione meccanica la quale ha condotto ai robot che vengono sviluppati oggi.

La maggior parte di questi sviluppi tecnici fu passata sotto silenzio dai più colti fra gli antichi greci e romani. Anche in tempi più vicini a noi gli studiosi dell'antichità classica non hanno dedicato molta attenzione ai testi sulle catapulte che sono sopravvissuti. L'importanza di questi testi si è rivelata appieno solo nel XX secolo, quando gli antichi trattati sulle catapulte cominciarono a essere decifrati da studiosi che disponevano di esperienza militare e di una formazione tecnica. Lo studio sperimentale di vere armi costruite sulla base di questi testi subito dopo l'inizio del secolo dall'ufficiale di artiglieria tedesco Erwin Schramm stimolò una linea di ricerche che è culminata in anni recenti con le opere definitive dello storico britannico E. W. Marsden. Questa letteratura chiarisce che i greci erano ben lontani dal disprezzare l'osservazione esatta ed esperimenti rigorosi come si suppone solitamente. Benché Platone possa avere disprezzato l'incapacità degli oggetti del mondo reale di adeguare le loro misure alle dimensioni ideali della geometria e benché Aristotele possa aver fondato in gran parte la sua biologia su descrizioni meramente verbali di specie animali, all'interno della comunità degli antichi ingegneri meccanici furono sviluppati metodi di studio della natura di importanza molto maggiore per il futuro.

La sostituzione dell'arco flessibile con l'elasticità di torsione dette un grande impulso all'ingegneria delle catapulte. Questo progresso fu compiuto circa mezzo secolo dopo l'invenzione della catapulta e qualche indizio suggerisce che Filippo di Macedonia, il padre di Alessandro Magno, fu il sovrano che sovvenzionò questa fase di ricerca e di sviluppo. Quel che è certo è che catapulte potentissime apparvero per la prima volta nel corso delle campagne di Alessandro. In effetti c'è ragione di associare l'ascesa di grandi imperi con l'avvento della catapulta.

Le catapulte che utilizzavano la potenza di un arco flessibile erano atte solo al lancio di frecce o pietre relativamente piccole (le pietre richiedevano una corda d'arco con una sorta di borsa o tasca nel suo punto di mezzo). Le frecce potevano avere le dimensioni di un piccolo giavellotto e le pietre dovevano essere abbastanza piccole da poter essere lanciate dalla spalla. Una volta perfezionato il principio della torsione, divenne possibile lanciare pietre pesanti fino a 78 chilogrammi. In effetti l'ingegnere romano Vitruvio ci dà le dimensioni di catapulte che potevano lanciare pietre fino a 162 chilogrammi di peso, anche se macchine così gigantesche non furono probabilmente mai costruite. Le macchine più comuni scagliàvano sassi di peso compreso fra 13 e 26 chilogrammi. Apparvero ora macchine in grado di lanciare frecce lunghe fino a quattro metri. Ma anche quando le catapulte lanciavano missili (nel senso originario di qualsiasi arma da getto) di scala più umana, l'arciere o il fromboliere veniva a essere completamente soverchiato. Le fonti dicono che una catapulta, che lanciava frecce di dimensioni normali (circa 70 centimetri) aveva una gittata massima di circa 640 metri, e c'è ragione di credere che si tratti di un'indicazione attendibile. La distanza massima raggiungibile da un arciere era di circa 450 metri.

Le catapulte erano in grado di scagliare tali proiettili con una precisone considerevole. Esse potevano facilmente concentrare i loro colpi su un singolo punto, colpendolo ripetutamente ed eliminando così i merli di protezione sulle mura di una città o staccando la corazza di una torre

In questa antica catapulta per frecce, fondata su un progetto elaborato da tecnici alle dipendenze di Dionisio il Vecchio, tiranno di Siracusa nel IV secolo a.C., un arco flessibile era montato all'estremità di una lunga struttura in legno comprendente una guida a coda di rondine. La slitta mobile, che trasportava la corda dell'arco grazie a un dispositivo di aggancio e scatto, era tirata verso la parte posteriore del fusto, contro la resistenza dell'arco, da un verricello. Una guida con denti di arresto solidale al fusto impediva il movimento in avanti della slitta, che veniva trattenuta da un nottolino. Pare che l'elemento che collegava la catapulta al piedistallo fosse una versione antica del giunto universale. Quanto all'arco, era costituito probabilmente da tre diversi materiali incollati assieme: una parte mediana in legno, uno strato frontale formato da tendini di animali e uno strato posteriore di corno. Poiché i tendini oppongono una forte resistenza alla tensione e il corno alla compressione, un tale arco doveva essere molto più potente di un arco comune ottenuto da un singolo pezzo di legno. La freccia è lunga circa 2 metri.

MOLLA DI TORSIONE

BRACCIO D'ARCO RIGIDO

GIUNTO UNIVERSALE

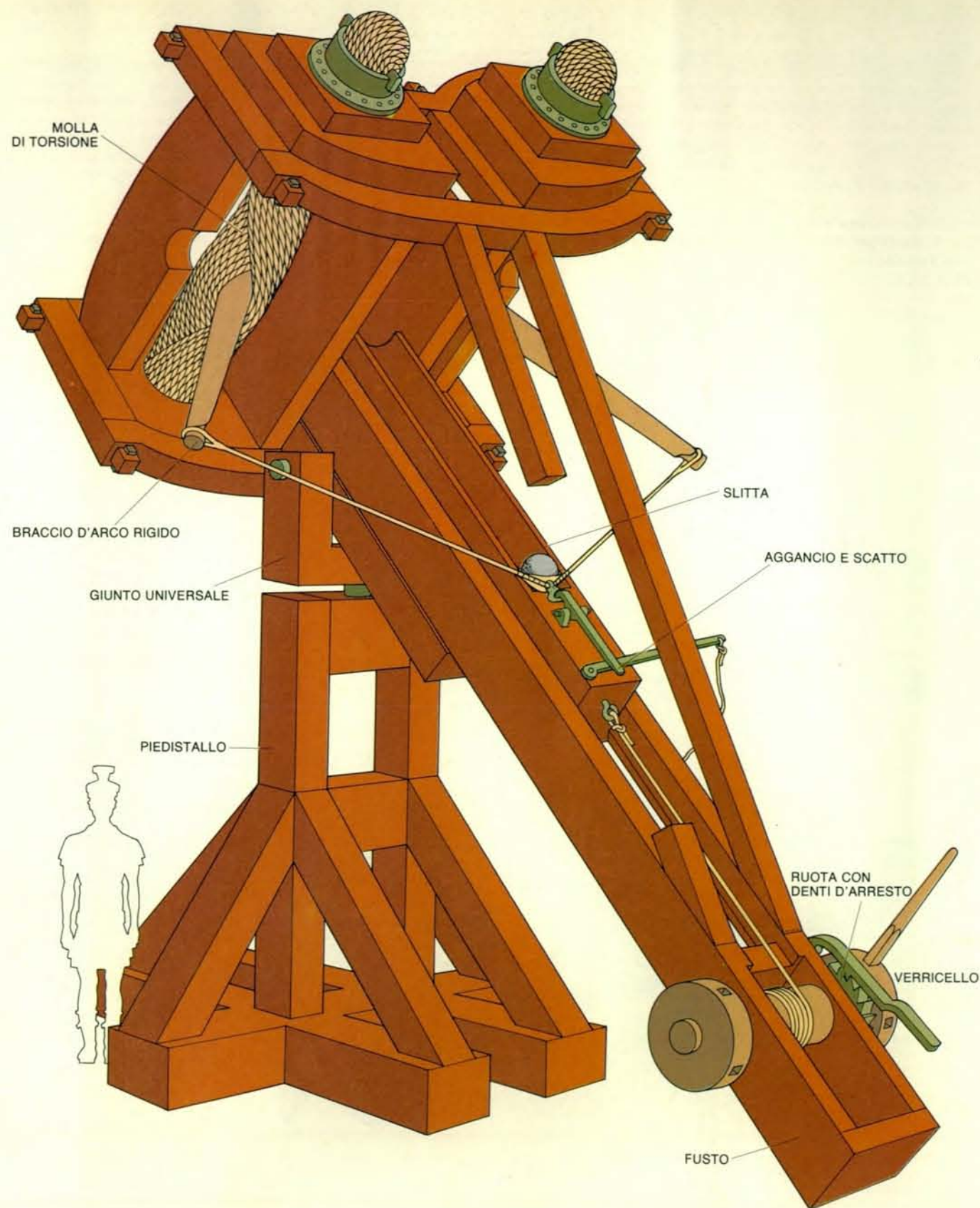
PIEDISTALLO

SLITTA

AGGANCIO E SCATTO

RUOTA CON DENTI D'ARRESTO

VERRICELLO

FUSTO

L'introduzione delle molle di torsione consentì agli antichi ingegneri costruttori di catapulte di progettare armi di dimensioni molto maggiori, come questa versione di catapulta o balista lanciasassi romana, la quale lanciava una pietra di un talento, ossia di 26 chilogrammi. (Una vista frontale di una macchina simile appare in copertina.) Una sorta di borsa o navetta al centro della corda contiene la pietra e un anello fissato alla parte posteriore della borsa è agganciato al dispositivo di bloccaggio. Le rondelle all'estremità dei fasci di torsione in questo particolare modello potevano essere fatte ruotare e venire poi fissate per adattare opportunamente la tensione prima di effettuare il tiro.

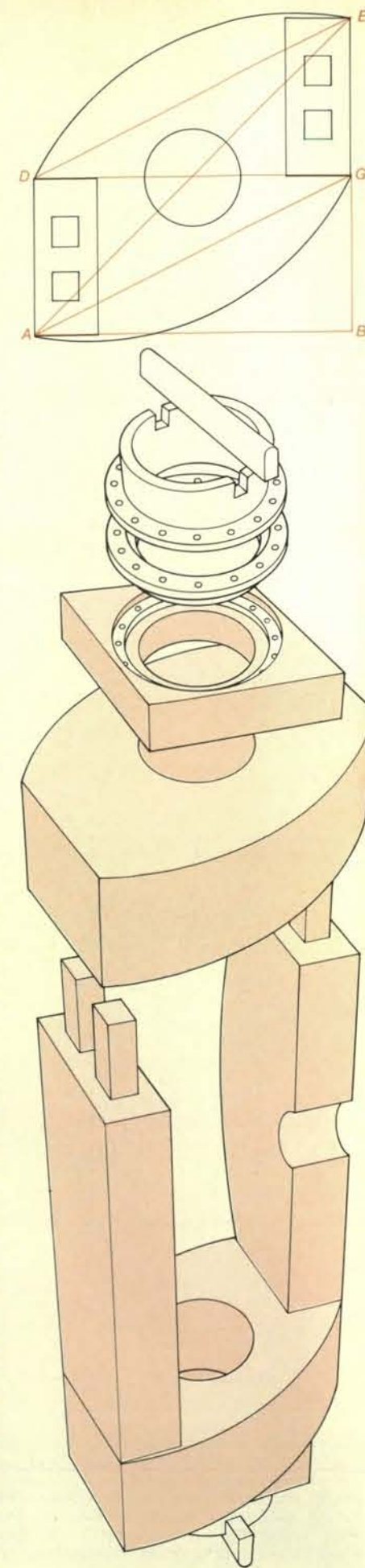
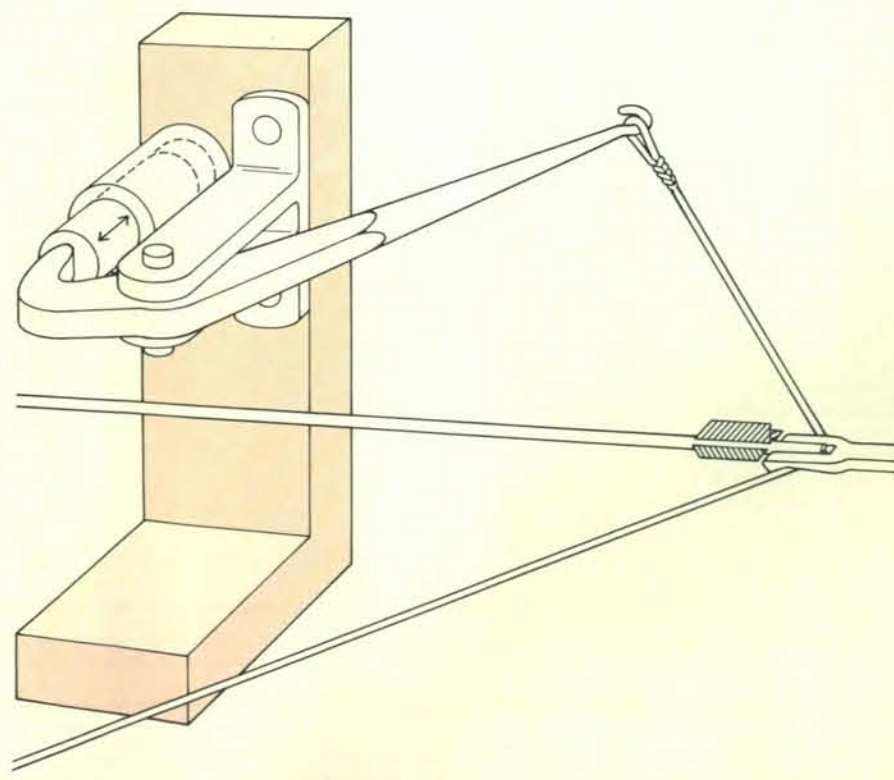
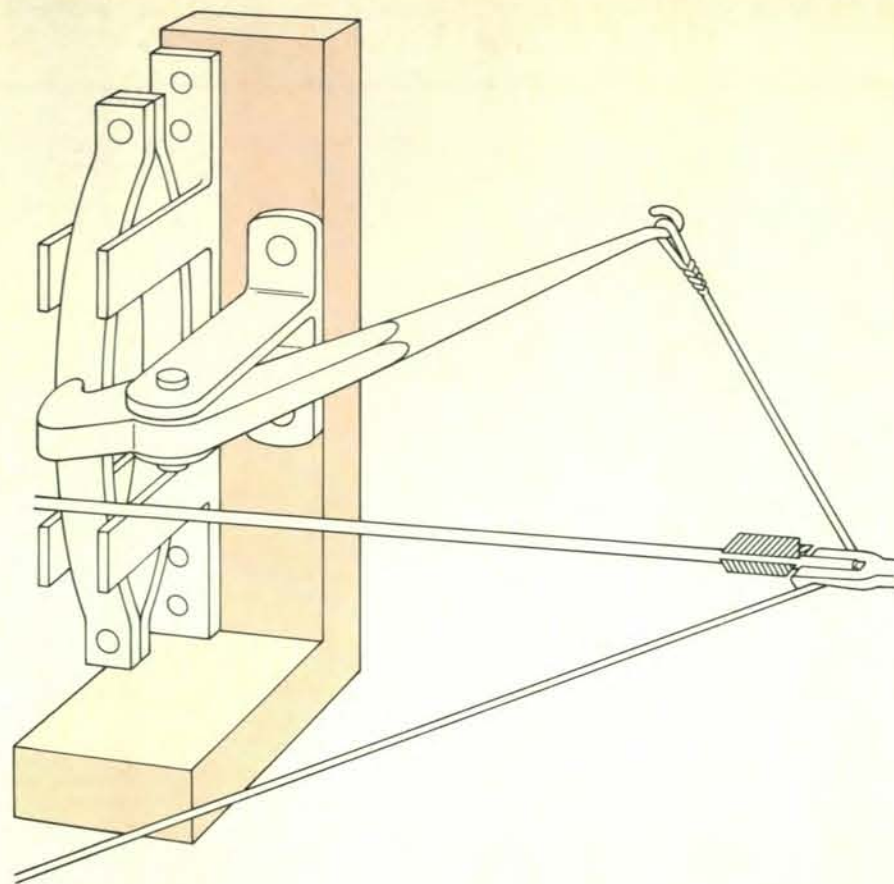
d'assedio mobile. Era possibile puntare le catapulte durante il giorno, quando si poteva osservare il punto di caduta dei missili, e poi terrorizzare i nemici tirando a intervalli nello stesso punto di notte. All'assedio di Avaricum (l'attuale Bourges), nel 52 a.C., Giulio Cesare osservò che le sue catapulte non avevano alcuna difficoltà ad abbattere uno dopo l'altro i guerrieri galli man mano che salivano in una postazione molto esposta che era di vitale importanza ai fini dell'attacco romano. Tutti questi particolari fanno pensare a un alto grado di precisione intrinseca di queste macchine. (In effetti, quando una di queste catapulte ricostruite da Schramm fu provata in presenza del Kaiser, si dice abbia spaccato in due con una freccia un'altra freccia andata precedentemente a bersaglio, nel migliore stile di Robin Hood.) È chiaro che la meccanizzazione aveva oltrepassato di molto le capacità degli arcieri umani.

All'assedio di Gerusalemme, nel 63 a.C., Giuseppe, il comandante delle forze ebraiche che difendevano la città, riferì che una palla lanciata da una catapulta romana aveva mozzato completamente la testa a un amico che stava accanto a lui sulle mura. Anche a distanze prossime a 400 metri una di queste palle poteva, a quanto pare, sfondare varie file di soldati prima di fermarsi. Nello stesso assedio, secondo Giuseppe, una donna gravida fu uccisa da una palla e il feto fu scagliato a 30 metri di distanza. Le frecce, lunghe e pesanti, erano altrettanto efficaci. Truppe in avanzata potevano essere letteralmente inchiodate al suolo dalla discesa di queste frecce. Di fatto la parola catapulta deriva dalla potenza di penetrazione di questa macchina. I fanti greci armati alla leggera portavano uno scudo chiamato *pelte*, e il prefisso *kata* denota un moto verso il basso. Una catapulta è quindi una macchina in grado di perforare completamente uno scudo verso il basso. All'assedio di Gaza, nel 332 a.C., Alessandro fu ferito al collo da una freccia di catapulta che gli aveva trapassato sia lo scudo sia la corazza. È noto che le macchine di Archimede suscitarono terrore all'assedio di Siracusa da parte dei romani dal 213 al 211 a.C. Una reazione tipica dell'epoca fu quella del generale spartano Archidamo, che, osservando il lancio di una catapulta, esclamò profeticamente: «Oh Ercole, il valore marziale umano non serve più!»

In definitiva, le catapulte incisero in misura significativa sulla conduzione della guerra e quindi sull'equilibrio di politica e società. A grandi linee, si può dire che spostarono il vantaggio dalla parte degli assedianti. Fino alla comparsa delle catapulte questi si erano trovati quasi sempre in una posizione di grande svantaggio. La guerra di Troia durò verosimilmente dieci anni e la lotta fra Atene e Sparta si protrasse certamente per un quarto di secolo, nonostante la grande superiorità dell'esercito spartano. A quei tempi la maggior parte delle città che cedevano agli assedianti venivano per lo più consegnate al nemico da traditori.

Con l'introduzione delle catapulte, e di altre macchine da guerra che stavano entrando in uso in Occidente, gli assedi divennero più efficaci. Dapprima Dionisio di Siracusa, e poi i macedoni Filippo e Alessandro, usarono alte torri da assedio mobili in grado di sovrastare le mura di una città e sottoporre i merli delle stesse a un bombardamento distruttivo, mentre enormi arieti, spinti fin da 1800 uomini, percuotevano le mura dal basso. Le catapulte non potevano certo competere in potenza con tali arieti, ma potevano abbattere mura che non erano state costruite a regola d'arte. Filone di Bisanzio, in un manuale di balistica composto attorno al 200 a.C., scrisse che un muro doveva avere uno spessore di almeno 4,62 metri per resistere alle pietre lanciate dalle catapulte e che era una buona idea cercare di impedire alle catapulte lanciasassi di avvicinarsi a meno di 150 metri per mezzo di fossati e di altri ostacoli.

Ma anche quando le mura erano ben costruite, i merli sporgenti al di sopra di esse restavano vulnerabili. Poiché i merli dovevano essere di spessore limitato allo scopo di consentire una buona visuale,

Le dimensioni del pannello che costituiva il pezzo superiore di uno degli alloggiamenti delle molle di torsione in una grande catapulta per il lancio di pietre venivano specificate dai progettisti in funzione delle dimensioni dei lati verticali dell'alloggiamento, i quali erano determinati a loro volta dal diametro del fascio di corde che formavano la molla di torsione. Lo spessore del pannello superiore non si conosce con sicurezza, ma probabilmente esso era press'a poco uguale al diametro del fascio di corde. Pare che i costruttori di catapulte procedessero disegnando dapprima un rettangolo con un lato uguale alla profondità dell'elemento verticale dell'alloggiamento e l'altro lato press'a poco doppio. Tracciavano poi la diagonale del rettangolo, da *A* a *G* e a partire da *D* tracciavano una linea parallela alla diagonale. La linea *BG* veniva poi estesa fino a intercettare la parallela da *D* nel punto *E*. Determinato in tal modo il punto centrale del parallelogramma *ADEG*, con centro in tale punto veniva tracciato un cerchio di diametro uguale a quello del fascio di corde. (Nel pezzo finito, il cerchio delimitava in effetti un foro che alloggiava il fascio di corde.) Venivano poi tracciati gli archi *DE* e *AG*, di raggio pari a tre volte il diametro del fascio di corde. Infine le mortase per i tenoni venivano centrate in modo appropriato nei due lati rettilinei del pezzo e scavate in esso per una profondità pari a due terzi del suo spessore. Gli ingegneri avevano portato i loro procedimenti di progettazione al punto di includere nelle loro istruzioni per la costruzione di catapulte metodi di trasformazione automatica della scala. Una volta scelti il sito e la destinazione dell'arma, si poteva stabilire la grandezza dei proiettili che si volevano lanciare. Una volta specificato questo elemento, la formula della catapulta forniva le dimensioni del fascio di torsione di cui c'era bisogno. Conoscendo il diametro del fascio, i manuali di costruzione, che si avvalevano dei risultati di decenni di sperimentazione, fornivano le dimensioni delle parti principali della macchina in multipli del diametro del fascio di corde. Nella costruzione reale le parti critiche della struttura in legno venivano rinforzate con elementi in ferro.

Due progetti alternativi per catapulte furono introdotti da Ctesibio di Alessandria alla metà del III secolo a. C. per sostituire il tipo con molle di torsione. In entrambi i progetti bracci d'arco rigidi ruotavano attorno a un perno in prossimità dell'estremo interno del braccio. Quando l'arco veniva teso, gli estremi interni dei due bracci si muovevano in modo tale da premere o su molle di bronzo *(in alto,* nel modello chiamato *chalcotonon)* o su pistoni che scorrevano all'interno di cilindri a perfetta tenuta d'aria *(in basso,* nel modello chiamato *aerotonon).* Nessuno dei due progetti fornì una forza confrontabile con quella dell'arco di torsione e i miglioramenti rimasero sulla carta.

era facile abbatterli con palle di pietra, non lasciando sulle mura alcun riparo per i difensori. Neutralizzati i tiri di reazione dei difensori, gli arieti e gli uomini che attaccavano le mura dal basso potevano lavorare con maggiore tranquillità. La catapulta ebbe un ruolo importante nel rendere la vita urbana nel IV secolo a.C. significativamente più precaria. Durante i suoi primi cinque anni di regno, Alessandro si impadronì di cinque grandi città e di molte città minori. Un brano della *Politica* di Aristotele (che fu maestro di Alessandro) riflette il mutamento. Un'urbanistica razionale, con strade rettilinee che si intersecavano a formare blocchi urbani rettangolari, era stata da poco popolarizzata in Grecia dall'architetto Ippodamo. Aristotele obiettò che almeno una parte di ogni città doveva conservare la disposizione casuale propria dei tempi anteriori per rendere più difficile agli invasori la penetrazione in essa. Inoltre, egli scrisse, «s'ha da ritenere che la difesa solidissima delle mura sia la protezione militare più adatta, tanto più con i moderni ritrovati in fatto di missili e di macchine d'assedio estremamente precise» (1330b-1331a; trad. R. Laurenti, Laterza, Bari, 1973).

Ma il consiglio di Aristotele non fu sufficiente. La piccola città-stato democratica stava morendo per opera della nuova tecnologia. Quando le sue mura divennero vulnerabili, essa fu inghiottita dai grandi imperi assolutistici come quello di Alessandro. La guerra campale democratica era fondata su uomini armati di aste, disposti spalla a spalla su una linea. Ogni guerriero svolgeva una funzione standardizzata e poteva essere facilmente sostituito in essa da un altro. Quest'uguaglianza sul campo di battaglia, assieme all'analoga uguaglianza dei vogatori nella flotta da battaglia, si era dimostrata un forte elemento a sostegno del sistema urbano democratico.

Con l'avvento di macchine militari specializzate, l'uguaglianza dell'armamento andò perduta. Speciali abilità matematiche e tecniche si richiedevano per la costruzione e la manutenzione di una catapulta e i rischi che si correvano nel servire una catapulta erano inferiori a quelli del combattimento in prima linea. In conseguenza di questa situazione cominciarono a emergere gerarchie di specialisti con funzioni e prerogative particolari. Nel corso del tempo un ordinamento politico che era appropriato a un livello di sviluppo tecnologico in cui la forza muscolare umana svolgeva una parte significativa risultò insufficiente quando i singoli cittadini-soldati cominciarono a perdere la loro posizione di preminenza a vantaggio delle macchine.

Nei nuovi ordinamenti politici l'antico ingegnere venne a trovarsi in una situazione favorevole. In passato il suo status non era stato elevato. Fra tutti gli dèi dell'Olimpo soltanto Efesto, che lavorava alla fucina e faceva ingegnosi lavori in metallo (compresi automi), veniva ritratto come sporco, deforme e zoppo. Lo scultore Fidia, che svolse un ruolo di supervisione nella costruzione dell'Acropoli sotto Pericle, fu accusato di sacrilegio per aver osato inserire un suo ritratto nella sua opera. In seguito, però, per gli artisti e artigiani la situazione migliorò. Non conosciamo il nome degli inventori al servizio di Dionisio, mentre ci sono pervenuti i nomi di tre fra coloro che lavoravano alla corte di Filippo e di Alessandro: Polidia, Diade e Caria. Nel secolo seguente i costruttori di catapulte divennero celebri e oggi sopravvivono i nomi e, almeno in parte, le opere di autori posteriori, come Erone, Filone, Bitone, Vitruvio, Ctesibio e altri. La tradizione culmina con Archimede, che era famoso al suo tempo più per le sue macchine da guerra che per le sue opere matematiche. Le recenti ricerche di A. G. Drachmann, dell'Università di Copenaghen, e di Derek J. de Solla Price della Yale University fanno ritenere sempre più inattendibile la tesi sostenuta da Plutarco che Archimede disprezzasse l'ingegneria. In epoca romana chi era esperto nel campo delle catapulte disponeva di un atout importante che poteva consentirgli anche di sopravvivere a epurazioni politiche.

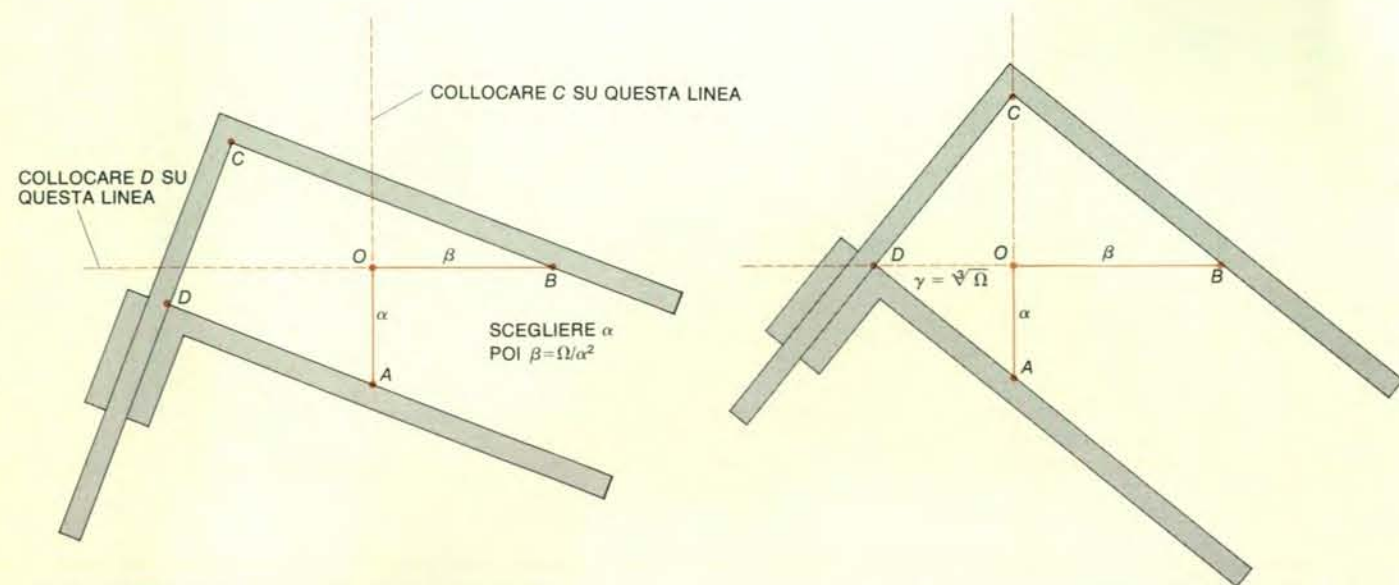
Questa promozione sociale dell'ingegneria si fondava su una forte domanda di catapulte. Esse divennero parte integrante di ogni fortezza e di ogni forza d'assedio moderne e gradualmente cominciarono a essere usate nella più mobile conduzione della guerra sul campo di battaglia. In mare le catapulte potrebbero avere avuto una parte nella corsa agli armamenti navali che condusse dalla trireme, con i suoi tre banchi di rematori, a navi enormi che contavano sino a 40 banchi di rematori. Evidentemente l'assunto che stava alla base di questi sviluppi era che il tiro con le catapulte poteva decimare le forze imbarcate sulle navi nemiche quando tali navi erano ancora troppo lontane per l'abbordaggio e lo speronamento. Quanto più grande era una nave, tante più catapulte poteva trasportare e tanto più stabile era la piattaforma di lancio. Secondo questa interpretazione, la catapulta avrebbe quindi soppiantato il com-

La catapulta più complessa inventata nell'antichità era un'arma a ripetizione progettata da Dionisio di Alessandria, attivo nell'arsenale di Rodi. Come si vede in questo particolare, le frecce scendevano nella scanalatura di lancio da una sorta di serbatoio o caricatore grazie alla semplice forza di gravità. Le frecce venivano fatte scendere una a una dalla rotazione di un tamburo di alimentazione scanalato in modo da accogliere una freccia per volta. La rotazione del tamburo era controllata da una camma elicoidale scanalata sulla superficie, nella quale era impegnato un dente montato sulla slitta. Il moto della slitta era prodotto a sua volta da due catene a maglie piane, una su ciascun lato della macchina. Secondo il testo pervenutoci che descrive la catapulta a ripetizione, le catene ingranavano su rocchetti in forma di prismi pentagonali. A giudizio dell'autore questi prismi dovevano funzionare come una sorta di ruota dentata inversa; in altri termini l'impegno della catena sui rocchetti pentagonali avveniva per mezzo di denti che entravano negli appositi incavi del rocchetto. Il prisma posteriore veniva fatto ruotare per mezzo di un verricello e il gancio per il bloccaggio della corda dell'arco veniva impegnato e liberato a tempo debito da spine montate in due posizioni estreme sul fusto dell'arma, passate le quali la slitta si muoveva. Perciò, invertendo la rotazione del verricello, la catapulta poteva automaticamente effettuare il lancio di tutte le frecce fino a quando il serbatoio non era completamente svuotato.

Un effetto dello sviluppo delle catapulte fu quello di costringere gli abitanti delle città a costruire mura più alte allo scopo di allungare la gittata delle loro macchine. Raddoppiando, per esempio, l'altezza delle mura non si aumentava la gittata in proporzione, ma si dava comunque ai difensori un vantaggio maggiore. Questo sviluppo nella storia delle fortificazioni suggerisce che le catapulte fossero considerate una minaccia maggiore di quella degli arieti, poiché mura più alte erano meno stabili quando venivano attaccate dal basso. Le catapulte difensive venivano usate probabilmente tenendone il fusto in posizione più o meno orizzontale per ottenere una maggiore precisione, come si può vedere in questo diagramma schematico. Raddoppiando l'altezza *(h)* si aumentava la gittata di una catapulta *(L)* della radice quadrata di 2.

battimento corpo a corpo in mare, svolgendo la parte che sarebbe stata assunta dai cannoni 2000 anni dopo. Infine, nuove tattiche di battaglia, di navi corazzate dette catafratte e gli sforzi dei romani per dominare l'intero Mediterraneo ebbero l'effetto di ridurre ancora una volta le dimensioni delle navi da battaglia.

L'importanza della catapulta condusse a sforzi per migliorarne le prestazioni anche oltre i limiti già raggiunti grazie all'introduzione dell'arco a torsione. L'ingegnere Ctesibio, per esempio, attivo ad Alessandria alla metà del III secolo a. C., tentò di eliminare le corde di capelli e di tendini, che erano soggette a rompersi, a marcire e a variare di tensione in seguito all'umidità e allo stiramento. Entrambi i progetti alternativi da lui proposti comprendevano bracci rigidi che ruotavano su un asse in prossimità del loro estremo più interno; questi bracci erano incurvati in modo tale che, quando l'arco veniva teso, premevano o su molle di bronzo martellato o su pistoni che scorrevano all'interno di cilindri a perfetta tenuta d'aria *(si veda l'illustrazione nella pagina a fronte)*. Ma né la compressione delle molle di bronzo (le quali sono ovviamente inferiori sotto la maggior parte dei punti di vista, a molle di acciaio) né la compressione della piccola quantità d'aria che può essere contenuta nei cilindri potevano fornire una forza paragonabile a quella dell'arco a torsione. (Nella sperimentazione di queste idee, Ctesibio scoprì che dal cilindro, assieme al pistone da lui forzato in esso con un martello, usciva del «fuoco». La fiamma o il fumo erano forse prodotti dall'accensione della colla da carpentiere da lui usata per ottenere una perfetta tenuta. Se l'accensione era causata dal calore di compressione dell'aria, Ctesibio può essere considerato lo scopritore dell'effetto diesel.)

Press'a poco nello stesso periodo, Dionisio di Alessandria sviluppò quella che fu forse la macchina più notevole di questo genere: una catapulta a ripetizione *(si veda l'illustrazione alla pagina precedente)*. Le frecce venivano alloggiate in un caricatore o serbatoio verticale alimentato a gravità e quindi trasferite una per volta nella scanalatura di tiro per opera di un tamburo rotante il cui moto era controllato da un sistema a guida eccentrica azionato dalla slitta. Non si conosce alcun esempio anteriore di un tale sistema, e nessun esempio altrettanto complesso è noto al Cinquecento. Un singolo moto del verricello controllava il tamburo, la slitta, il meccanismo di agganciamento e di scatto, cosicché era sufficiente far ruotare il verricello avanti e indietro per mantenere in azione la macchina sino allo svuotamento del caricatore. Qui fece per la prima volta la sua comparsa la catena a maglie piane, spesso attribuita a Leonardo. Le maglie della catena avevano presumibilmente sporgenze che andavano a inserirsi negli interstizi di un ingranaggio inverso; in altri termini, i denti erano interni, non esterni, un po' come avviene in una moderna sega a catena. (Questa interpretazione si fonda in parte su particolari contenuti nel testo che si è conservato e in parte sulle necessità meccaniche della situazione.)

L'estrattore di radice cubica, un semplice dispositivo meccanico inventato da un ignoto geometra greco del III o IV secolo a.C., rendeva possibile progettare una catapulta lanciasassi a una determinata scala risolvendo una formula, la quale stabiliva che il diametro (in dita) del fascio di corde era uguale a 1,1 volte la radice cubica di 100 volte il peso dei proiettili di pietra che si volevano lanciare espresso in mine. Per trovare la radice cubica di un certo valore Ω, per esempio, si sceglieva prima un segmento lineare Ω e poi si otteneva il valore ß tale che $\beta = \Omega / \alpha^2$. Le linee α e ß venivano poi disegnate ad angoli retti *(a sinistra)*. Facendo scorrere la ganascia mobile del dispositivo, lo si poteva allineare in modo che il punto *C* venisse a trovarsi sul prolungamento verticale della linea α e il punto *D* sul prolungamento orizzontale della linea ß *(a destra)*. La soluzione era data allora da γ, un valore uguale alla distanza *OD*. Supponiamo, per esempio, che γ fosse uguale alla radice cubica di 100. Se si dovesse scegliere α uguale a 4, ß dovrebbe essere uguale a 100/16, ossia 6,25. L'estrattore di radice cubica, dopo essere stato allineato in maniera opportuna, fornirebbe per γ il risultato approssimativamente corretto, ossia 4,64.

La catapulta a ripetizione non riuscì a sostituire la catapulta normale. La facilità e rapidità di operazione andavano a scapito della gittata, che era sensibilmente inferiore. La sua stessa precisione, inoltre, finì paradossalmente col costituire uno svantaggio. La macchina concentrava i suoi tiri alla sua distanza massima (circa 200 metri) in uno spazio così ristretto che non valeva la pena di tirare a quella distanza neppure su un piccolo gruppetto di uomini. (Fu un modello di una di queste catapulte a ripetizione a spaccare in due una freccia nell'esibizione di tiro di Schramm per il Kaiser.) I comandanti temevano inoltre che essa comportasse uno spreco di munizioni: la stessa lagnanza che fu sollevata contro l'invenzione dei fucili a ripetizione due millenni dopo.

Un'altra ragione dell'insucesso di queste interessanti variazioni può essere vista nel grado di sofisticazione degli sforzi tecnici applicati nel frattempo alla catapulta comune. Il successo della catapulta rese imperativa l'esigenza di raggiungere gittate almeno pari a quelle raggiunte dai nemici. Fu allora necessario adeguare la quantità di fibra elastica al peso dei proiettili. Probabilmente la progettazione non mirò tanto a raggiungere la gittata massima in assoluto quanto a pervenire al punto oltre il quale i costi aumentavano troppo rapidamente e diminuivano la maneggevolezza e la precisione (quest'ultima in conseguenza di fattori balistici lungo la traiettoria).

Uno dei passi avanti cruciali nel progettare l'utilizzazione delle molle di torsione consistette nello stabilire un rapporto fra il diametro e la lunghezza del fascio cilindrico di corde elastiche. Se le corde erano troppe corte, sviluppavano un elevato attrito interno e potevano non avere un allungamento elastico sufficiente per evitare la rottura quando i bracci dell'arco venivano tirati all'indietro. Se erano troppo lunghe, una parte della loro elasticità rimaneva inutilizzata pur tirando i bracci all'indietro nei limiti consentiti dalla struttura della macchina. Tutte le descrizioni di catapulte che si sono conservate fanno pensare che in realtà venisse raggiunta una configurazione cilindrica ottimale e che da essa non ci si allontanasse tranne che in circostanze particolari, come nel caso delle macchine esclusivamente a breve gittata costruite da Archimede a Siracusa.

Questa ottimizzazione del fascio di corde fu raggiunta attorno al 270 a.C., forse dal gruppo di ingegneri greci che lavoravano per la dinastia tolemaica in Egitto. Ivi e a Rodi gli esperimenti dei ricercatori sulle catapulte furono, secondo Filone, «massicciamente sovvenzionati perché essi avevano re ambiziosi che

promuovevano la loro abilità artigianale».
I risultati furono compendiati in due formule. Per una catapulta a frecce il diametro del fascio di corde fu stabilito semplicemente in 1/9 della lunghezza della freccia. La formula, più complessa, di una catapulta lanciasassi stabiliva, in termini moderni, che il diametro del fascio di corde in dita (un dito vale press'a poco 19,8 millimetri) è uguale a 1,1 volte la radice cubica di 100 volte il peso della palla in mine (una mina è pari a 437 grammi circa).

La formula della catapulta lanciasassi presenta due caratteristiche notevoli. Innanzitutto dà una soluzione reale ed esatta per una progettazione ottimale. Per rendercene conto, supponiamo dapprima (supposizione in effetti ragionevole) che i costruttori di catapulte desiderassero dare alle loro macchine prestazioni massime. Essi dovevano a tal fine massimizzare l'energia cinetica dei loro proietti, e quindi l'energia potenziale immagazzinata nelle molle a torsione. La moderna teoria dell'elasticità applicata al disegno di queste molle ci dice che l'energia immagazzinata disponibile per il lancio sarà proporzionale alla quantità della tensione iniziale data al fascio nel suo montaggio nella catapulta più la tensione addizionale causata dal preattorcigliamento del fascio, al quadrato dell'angolo indicante l'ulteriore attorcigliamento conferito tirando all'indietro il braccio dell'arco e al cubo del diametro della sezione del fascio. L'elevazione al cubo del diametro del fascio significa che, per esprimere il diametro del fascio elastico nei termini della massa del proietto, si doveva estrarre una radice cubica.

Un tardo sviluppo nella storia della tecnologia delle catapulte fu la cheiroballistra o carrobalista romana, una macchina su ruote relativamente piccola che risale al I secolo d.C. circa. La struttura in ferro le conferiva una mobilità sufficiente per consentire l'uso sul campo di battaglia e la struttura aperta facilitava la localizzazione di bersagli mobili. L'arma veniva puntata probabilmente allineando la punta della freccia con un traguardo, opportunamente elevato, montato sulla parte posteriore del fusto. Le relazioni matematiche fra la forza dell'arco, lo spostamento della slitta e l'angolo di elevazione erano tali che, usando una linea di mira che passava per il bersaglio, la punta della freccia e il traguardo posteriore, si poteva conseguire automaticamente la traiettoria corretta solo stimando la distanza e tirando indietro la slitta per il numero opportuno di denti d'arresto. La cheiroballistra fu poi soppiantata da una catapulta più semplice, a un solo braccio, chiamata onagro.

Si osservi che, per arrivare a questo risultato, si devono usare i concetti dell'energia cinetica e potenziale, i quali non furono combinati in una relazione significativa fino al Settecento, ossia fino all'opera di Eulero e Daniel Bernoulli. Si richiede inoltre la teoria dell'elasticità, la quale era stata iniziata circa mezzo secolo prima da Hooke e da Robert Boyle. Si devono impiegare inoltre i principi della balistica, i quali non furono chiariti sino all'opera di Bonaventura Cavalieri e di Galileo Galilei, nella prima metà del Seicento. Che gli antichi ingegneri delle catapulte riuscissero a pervenire a una formula che conserva la sua validità anche alla luce di questi sviluppi molto posteriori è un fatto davvero impressionante.

L'utilizzazione di una radice cubica costituisce il secondo carattere notevole della formula della catapulta lanciasassi, in quanto tale formula fu scritta in un periodo in cui i matematici greci non erano ancora in grado di padroneggiare appieno le equazioni di terzo grado. Attorno al 460 a.C. Ippocrate di Chio (da non confondersi col famoso medico) aveva stabilito che sarebbe stato possibile duplicare con precisione il volume di un cubo se si fossero trovate due linee che definissero due medi proporzionali fra due linee date. Nel secolo seguente si era fatto solo qualche passo avanti verso la soluzione del problema. Archita di Taranto ed Eudosso di Cnido avevano escogitato eleganti soluzioni teoriche, ma si trattava di soluzioni tridimensionali, molto scomode fisicamente e quasi di nessuna utilità nell'esecuzione di calcoli. La situazione era a questo punto quando fu introdotto l'arco a torsione.

La maggior parte di coloro che risolsero il problema della radice cubica avevano una connessione diretta e indiretta con le catapulte. Menecmo, secondo la tradizione, fu uno degli insegnanti di Alessandro Magno; perciò era presente nel tempo e nel luogo in cui l'arco a torsione venne per la prima volta alla ribalta. La sua soluzione implicava l'intersezione di sezioni coniche, un concetto che pare sia stato scoperto da lui. Purtroppo non ci sono elementi sufficienti che consentano di stabilire se egli fu condotto a considerare le coniche dal problema di estrarre radici cubiche necessarie per la progettazione di catapulte.

Dopo di lui si adoperò alla soluzione del problema delle radici cubiche Eratostene, che era nato a Cirene ma viveva ad Alessandria, la quale era allora un centro della ricerca sulle catapulte. Eratostene, che era amico di Archimede, affermò esplicitamente che la catapulta era la principale ragione pratica per affrontare problemi con radici cubiche. Possiamo supporre che Eratostene fosse interessato a problemi di ingegneria, dal momento che Archimede gli dedicò il suo libro *Sul metodo*. In quest'opera soluzioni approssimate di problemi matematici sono abbozzate inizialmente con un'impostazione pratica di tipo meccanico. Per esempio, talune sezioni di corpi vengono pesate per determinare il rapporto dei loro volumi. La soluzione di Eratostene si fondava su un dispositivo meccanico con parti scorrevoli, che doveva essere simile a quello raffigurato nell'illustrazione di pagina 92 in basso.

Il gruppo di uomini che si occuparono subito dopo di lui del problema della soluzione delle radici cubiche, fra cui Filone di Bisanzio, Archimede di Siracusa ed Erone di Alessandria, erano tutti famosi per i loro lavori sulle catapulte. È interessante osservare che la più grande catapulta lanciasassi di cui si abbia notizia, una macchina da tre talenti (78 chilogrammi), fu costruita da Archimede. Una macchina di questo «calibro» avrebbe richiesto un'estrazione di radice cubica, in quanto non ci sono radici naturali nella quantità desiderata. Nella costruzione di queste macchine a breve gittata Archimede dovette anche allontanarsi dalle proporzioni delle catapulte normali. L'efficacia delle catapulte da lui costruite attesta la sua abilità come ingegnere matematico.

Considerati come gruppo, questi antichi studiosi del problema della radice cubica stanno a sé rispetto all'orientamento principale della matematica greca. Anziché limitarsi all'uso della riga e del compasso, essi escogitarono semplici dispositivi meccanici che consentivano loro di generare sezioni coniche e anche curve di ordini superiori. Alcuni ausili di cui si valsero, come il semplice «regolo calcolatore» raffigurato nell'illustrazione a pagina 92 in basso, anticipano i compassi di proporzione e gli archipendoli usati dagli artiglieri verso la fine del Cinquecento. L'opera di questo gruppo fu però trascurata sino al Rinascimento, quando la matematica riprese a svilupparsi press'a poco a partire dal punto in cui essi l'avevano lasciata. *La géometrie* di Descartes, per esempio, si inizia con procedimenti e dispositivi molto simili ai loro.

Pare, perciò, che gli ingegneri che si occuparono di catapulte eseguissero esperimenti che li costrinsero a entrare in un campo fino ad allora ignorato dai procedimenti matematici tradizionali. È abbastanza facile ancora oggi adattare dati di terzo grado a una curva di secondo grado se i dati sono scadenti o l'investigatore è poco scrupoloso. Perciò dobbiamo avere una grande considerazione per questi antichi investigatori. Essi devono avere ripetuto molte volte le loro prove di tiro con le catapulte, devono avere annotato i dati delle osservazioni con estrema diligenza e devono averne interpretato i risultati con grande precisione. I passi introduttivi dei *Belopoeika* di Filone insistono molto sui procedimenti sperimentali e sui risultati ottenuti dagli antichi ingegneri costruttori di catapulte, e dal punto di vista della teoria dell'ingegneria moderna l'esattezza della sua esposizione pare confermata completamente.

Essendo pervenuti a un volume e a una configurazione ottimali per il fascio di corde che dovevano fornire l'elasticità di torsione, gli ingegneri delle catapulte continuarono i loro esperimenti fino a ottimizzare le dimensioni delle restanti parti costitutive delle loro macchine da getto. Se i bracci dell'arco erano troppo corti, la forza richiesta per il caricamento poteva essere eccessiva, lo spostamento della corda dell'arco troppo piccolo e le sue capacità di trasferimento di energia potevano risultarne diminuite. Se i bracci dell'arco erano troppo lunghi, ritardavano l'azione delle molle a causa della loro massa eccessiva, oppure rendevano la macchina troppo ingombrante. Stabilita la lunghezza dei bracci dell'arco, dal tragitto della corda dell'arco si potevano determinare la lunghezza della slitta e del fusto, e così via per il resto della macchina.

Infine gli ingegneri delle catapulte scrissero i loro testi in modo tale da fornire le dimensioni delle parti principali della macchina come multipli del diametro del fascio elastico. Una volta calcolato questo diametro per la grandezza del proiettile desiderato, il resto della macchina veniva trasformato alla scala appropriata; e questo attesta un livello di razionalità tecnica che non fu raggiunto di nuovo fino all'epoca della Rivoluzione Industriale.

L'ultimo perfezionamento importante nella progettazione delle catapulte risale all'epoca romana tarda, quando il ferro fu sostituito come materiale da costruzione al legno. Quest'innovazione rese possibile una riduzione nelle dimensioni, un aumento dei livelli di sollecitazione e una maggiore libertà di corsa per i bracci dell'arco. La catapulta poté ora essere montata su ruote e spostata con facilità da una parte all'altra del campo di battaglia in risposta alle mutevoli vicende della lotta. La nuova struttura aperta semplificò anche il puntamento, che nelle macchine anteriori in legno era stato assai ridotto, specialmente per bersagli mobili e vicini.

Questo tipo avanzato di catapulta - la carrobalista - venne troppo tardi per poter essere usato nel periodo di espansione della civiltà romana, ma svolse una sua funzione nello stabilizzare i confini dell'impero e nel contribuire a impedirne l'erosione. Man mano che l'Impero declinava, però, le capacità tecniche richieste per costruire e usare macchine così complesse pare diventassero sempre più rare. Una nuova macchina più semplice, detta onagro, con una sola fascia elastica e un solo braccio terminante con un cucchiaio, che veniva usata per scagliare pietre, si impose sempre più. Essa avrebbe fornito l'artiglieria pesante del Medioevo, fino alla comparsa del trabocco, la cui struttura era ancora più semplice e il cui funzionamento sfruttava la forza di gravità.

La progettazione scientifica di macchine complesse, con una deliberata coordinazione sperimentale delle dimensioni delle parti componenti, non ricomparve nella civiltà occidentale sino al Settecento. Nel periodo antico i mutamenti in cui ebbe parte la catapulta prefigurarono, in modi che non possono non colpirci, problemi riapparsi molto tempo dopo nei rapporti fra scienza e tecnologia da un lato e guerra e società dall'altro.

I delfini

Questi discendenti di mammiferi terrestri adattatisi all'ambiente acquatico hanno un cervello di grandi dimensioni, facilità di apprendimento e un ricco repertorio vocale; eppure mancano dati sicuri sul loro grado di intelligenza

di Bernd Würsig

I delfini godono fama di possedere un livello d'intelligenza superiore a quello di qualsiasi altro animale e persino uguale a quello dell'uomo, perché sono dotati di un cervello di grandi dimensioni, hanno una notevole prontezza nell'apprendere i giochi escogitati dagli ammaestratori e mostrano un ricco repertorio di segnali vocali. Sulla base delle osservazioni compiute sui delfini in molti habitat naturali, credo che, nonostante tutti gli sforzi, sia impossibile trovare una collocazione precisa per questi animali nella classifica dell'intelligenza animale. Allo stato attuale della tecnologia marina è semplicemente impossibile per gli osservatori riuscire a rimanere per periodi prolungati accanto a un animale che vive in pieno oceano e nuota velocemente per lunghi percorsi. Quando si conoscerà più a fondo il comportamento dei delfini, le domande che ci si pongono sull'intelligenza di questi mammiferi troveranno una risposta soddisfacente. Attualmente il massimo che si possa dire sui delfini è che sono animali gregari paragonabili nel comportamento individuale e sociale ai mammiferi terrestri che vivono in armenti e in branchi.

I delfini si sono evoluti almeno 50 milioni di anni fa da mammiferi terrestri che forse rassomigliavano agli ungulati artiodattili viventi attualmente, come i bovini e i suini. Dopo essersi cimentati con l'ambiente marino i delfini si adattarono progressivamente alla vita acquatica: la pelliccia ancestrale venne sostituita da uno spesso strato di grasso sottocutaneo, il corpo divenne liscio e affusolato, gli arti posteriori si ridussero a ossa rudimentali interne, soppiantati nella loro funzione propulsiva da una robusta coda, gli arti anteriori furono trasformati in natatoie; molte specie poi, evidentemente allo scopo di migliorare la capacità di governare e stabilizzare il nuoto, svilupparono una pinna dorsale.

I delfini, di cui sono state classificate più di 30 specie, appartengono al sottordine degli odontoceti, dell'ordine dei cetacei. (Le focene, assai simili ai delfini e perciò spesso confuse con questi, sono pure rappresentanti degli odontoceti: si distinguono dai delfini perché hanno un muso meno simile a un becco e denti compressi in senso laterale, a forma di vanga. Si dovrebbe osservare tuttavia che molti studiosi americani di mammiferi marini danno il nome di focene, *porpoise*, a tutti i piccoli cetacei odontoceti, senza considerare attentamente le caratteristiche fisiche di questi animali.) Alcune specie di delfini vivono in acque dolci, ma la stragrande maggioranza frequenta un habitat oceanico. Le specie d'acqua dolce viaggiano in piccoli gruppi o sono pressoché solitarie, mentre le specie oceaniche (come il delfino macchiato del Pacifico) possono riunirsi in gruppi di parecchie migliaia. Tali cifre ricordano le mandrie di bisonti dell'America settentrionale e dei vari erbivori che pascolano nelle pianure di Serengeti in Africa: si resta stupefatti di fronte alle rassomiglianze etologica ed ecologica tra i delfini e questi lontani parenti della terraferma.

In generale, i mammiferi a elevata organizzazione sociale possiedono un ricco repertorio di complicati segnali sociali e di comportamenti; per questo motivo possono interagire con altri membri del gruppo in modi molto complessi. Tra i vari esempi si possono citare i segnali d'aggressione che sono utili nello stabilire e mantenere le gerarchie di dominanza, i segnali di corteggiamento, i suoni d'avvertimento o gli spostamenti all'avvicinarsi di un aggressore potenziale e molti altri segnali che contribuiscono al buon funzionamento del gruppo e al benessere dei membri che lo costituiscono.

Lo sforzo di accumulare dati sul comportamento e sui sistemi sociali dei delfini è reso particolarmente difficile dal fatto che quasi tutti i canali di comunicazione di questi animali si stabiliscono sotto le onde. È estremamente difficile avvicinarsi a un gruppo di delfini stando in una barca e rimanere tra di loro per un periodo abbastanza lungo per cominciare a capire il sistema sociale di questi mammiferi. Tutto ciò che l'osservatore vede è un gruppo di pinne dorsali, quando gli animali vengono in superficie per respirare; poi li si perde di vista quando avanzano sott'acqua. Anche nei rari casi in cui l'acqua è calma e limpida, e i delfini si possono vedere per periodi di parecchi minuti, la vicinanza di una barca può disturbarli: perciò risulta difficile separare ciò che è naturale nel loro comportamento dalle azioni innaturali prodotte solo dalla reazione alla vicinanza di una barca.

A causa di questi e di altri problemi, quasi tutte le prime osservazioni venivano compiute un tempo su delfini in cattività. Benché la cattività costituisca una situazione scomoda per animali abituati a una vita libera, con scarsi limiti fisici, tuttavia parecchi etologi hanno potuto raccogliere utili dati sui delfini catturati semplicemente perché avevano agio di osservarli per periodi prolungati di tempo. La ricerca più nota in questo campo è quella compiuta da Margaret C. Tavolga di Marineland, in Florida, che potè ossservare un gruppo di 12 tursiopi, chiamati anche delfini dal naso a bottiglia *(Tursiops truncatus)* in una grande vasca per un periodo totale di cinque anni.

Tavolga trovò che il gruppo possedeva una precisa gerarchia di dominanza. L'unico maschio adulto, che era l'animale più grande del gruppo, era più aggressivo e più intrepido di qualsiasi altro membro femmina, maschio subadulto o giovane delfino. In generale gli animali più grossi erano dominanti sopra i più piccoli.

Gregory Bateson, dell'Oceanic Institute di Hawaii, ha ricavato dati analoghi: in un gruppo costituito da due delfini macchiati *(Stenella attenuata)* e da cinque esemplari di *Stenella longirostris*, esisteva una simile gerarchia di dominanza. Il maschio di maggiori dimensioni minacciava spesso gli altri delfini, colpendoli o mostrando loro i denti, mentre non veniva mai minacciato. Il delfino che occupava la seconda posizione gerarchica, anch'esso un maschio, minacciava gli animali di classe inferiore, e così via lungo tutta la linea gerarchica. Le osservazioni di Bateson, come quelle di altri ricercatori, dimostrano come la gerarchia dei delfini non sia così stretta come in alcuni altri mammiferi. Per esempio, il maschio di classe infima ha ancora la possibilità di accoppiarsi con una femmina senza essere sfidato dal maschio più grosso.

Le tre fasi del salto tipico di un delfino sono state riprese casualmente in questa fotografia scattata nel Mare di Cortés, al largo della costa messicana. Si tratta di tre delfini comuni *(Delphinus delphis)*. Il repertorio di salti, contorsioni, capriole eseguito dai delfini è ricco e vario.

Un branco di delfini della specie *Delphinus delphis* è stato fotografato nell'Atlantico mentre compie salti. I delfini sono mammiferi sociali che talvolta si aggregano in branchi di notevoli dimensioni, come in questo caso, ma più spesso si riuniscono in sottogruppi di una ventina di individui. Poiché sono animali che respirano aria, i salti servono a questo scopo. I delfini saltano molto anche quando giuocano o cacciano.

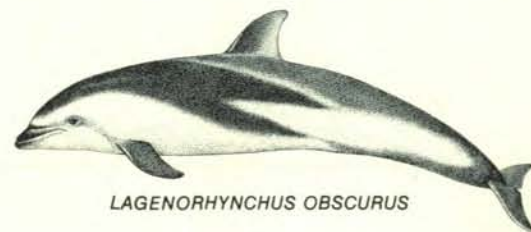
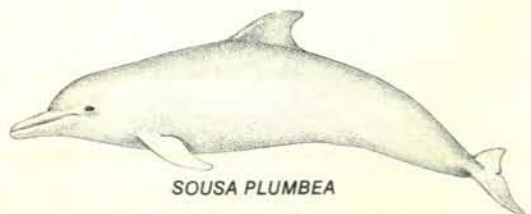
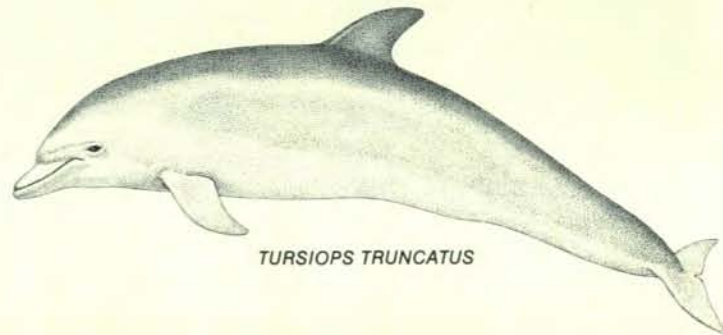
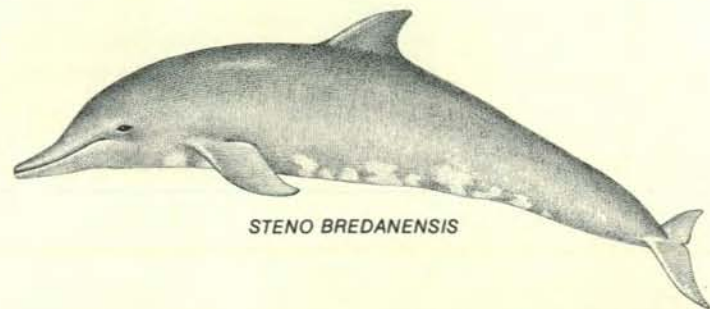
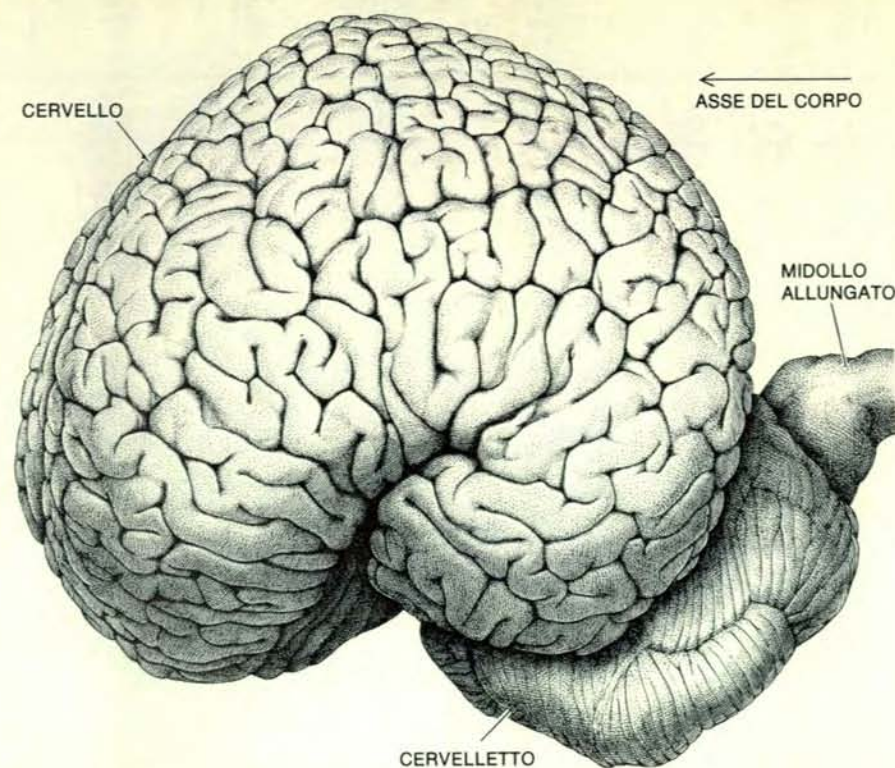
Poiché la gerarchia di dominanza non è necessariamente in relazione con le possibilità di avvicinare una femmina da parte di un maschio (come avviene per esempio nell'harem della foca elefantina), ci si domanda quale sia la sua funzione nelle popolazioni di delfini in libertà. Kenneth S. Norris e Thomas P. Dohl dell'Università della California, a Santa Cruz, hanno proposto l'ipotesi che la funzione possa essere quella di organizzare i membri del gruppo in modo da far fronte a diverse situazioni. Per esempio, i segnali di minaccia e gli inseguimenti che i delfini più grossi eseguono tenderebbero ad aggregare le femmine più piccole e i giovani al centro del gruppo, dove sarebbero più protetti da predatori potenziali come squali e orche. Norris afferma di aver osservato tale organizzazione nei delfini macchiati e in quelli della specie *Stenella longirostris*.

Affinché un sistema di tal genere possa evolversi è utile e forse necessario che gli animali che ne sono dotati presentino affinità genetiche tra loro. Secondo l'opinione di W. D. Hamilton dell'Imperial College of Science and Technology e di altri biologi, gli animali che presentano una stretta relazione di parentela dovrebbero cercare di proteggersi a vicenda, più di quanto proteggano i parenti più lontani, poiché i parenti più affini hanno in comune un numero maggiore di geni. Se una madre salvaguarda la sua prole facendola raggruppare all'interno del branco e mettendo se stessa in maggior pericolo, dimostra evidentemente un tipo di altruismo adattativo per il gruppo, perché permetterà di mettere in salvo una percentuale significativa dei propri geni.

Se i delfini riuniscono i membri del branco in questo modo, è probabile che almeno alcuni degli animali del gruppo abbiano rapporti di parentela. Si sono visti anche delfini che aiutavano un loro compagno, malato o sofferente, permettendogli di raggiungere la superficie per respirare, e che proteggevano un altro compagno dagli attacchi dei predatori e da altri pericoli. Questi modelli di comportamento sono stati spesso citati come prove evidenti di altruismo quasi umano e di grande intelligenza. Sembra più probabile che rappresentino una conseguenza di una tendenza evolutasi nell'aiutare gli individui geneticamente affini.

Purtroppo i gradi di consanguineità in una popolazione di delfini in libertà non sono noti. Solo di recente si è fatta strada una vaga idea sulla struttura sociale tra i delfini; la conoscenza è complicata dall'esistenza di molte specie diverse e di parecchie popolazioni ancor più separate nell'ambito delle singole specie. Dai mammiferi terrestri abbiamo imparato che il sistema sociale rappresenta in parte un adattamento all'habitat della popolazione. Per esempio, i lupi che vivono principalmente a spese dei cervi tendono a spostarsi in piccoli branchi, mentre quelli che vanno a caccia di alci si raggruppano in società più numerose e organizzate in modo più sofisticato. Questa differenza è evidentemente legata alla necessità di uno sforzo coordinato da parte di parecchi animali per abbattere una specie di maggiori dimensioni.

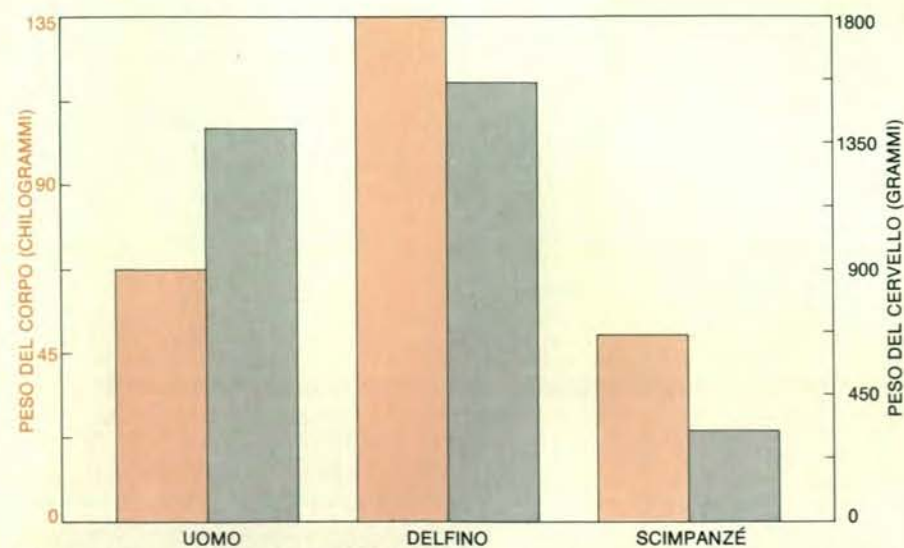
Recentemente sono stati compiuti alcuni studi che permettono utilissimi confronti tra le specie e le popolazioni di delfini. Norris e Dohl hanno studiato dalle scogliere affacciate al mare e da sott'acqua la *Stenella longirostris* delle Hawaii. Hanno scoperto che questi mammiferi, caratterizzati da elevata organizzazione sociale, si spostano in branchi composti in media da 25 individui. La struttura del branco varia durante il giorno in modo prevedibile. Al mattino i delfini si spostano lentamente e in gruppi serrati, in modo che gli animali quasi si toccano fra loro: in tal modo sembra che si riposino. Più tardi divengono sempre più vivaci, nuotano più velocemente, alcuni di loro spiccano salti nettamente al di sopra della superficie, compiendo rotazioni, capriole e tutti quegli altri giochi acrobatici per cui i delfini sono tanto apprezzati negli acquari.

A questo punto i branchi si sparpagliano, con gli animali che spesso tengono una distanza di una ventina di metri tra loro. Inoltre i gruppi tendono a congiungersi, in modo che si forma un branco ampliato di 50 e più membri, con tutti gli animali che si spostano nella medesima direzione. Quando si avvicina la notte, il branco migra parecchi chilometri al largo delle

Nell'illustrazione sono raffigurate cinque specie di delfini, per mettere in evidenza le diversità di forma e di distribuzione delle macchie. Sono il tursiope *(Tursiops truncatus)*, *Stenella longirostris*, il lagenorinco scuro dell'Atlantico meridionale *(Lagenorhynchus obscurus)*, il delfino degli Oceani Indiano e Pacifico *(Sousa plumbea)* e il delfino dai denti rugosi *(Steno bredanensis)*, disegnati in scala. Oltre a queste, sono note altre 25 specie di delfini. Le focene, assai simili ai delfini, hanno però il muso più tozzo e sono più corte.

L'illustrazione mostra il cervello di un delfino della specie *Turpiops truncatus* (tursiope troncato, o delfino dal naso a bottiglia). Il cervello di un delfino pesa in media circa 1,60 chilogrammi; quello di un uomo pesa in media 1,36 chilogrammi, ma è più grande rispetto al peso corporeo.

coste, entrando in acque profonde e iniziando a compiere immersioni profonde per andare a caccia di pesci, a parecchie centinaia di metri sotto la superficie. La popolazione di delfini si dispone in questo momento in molti gruppi largamente sparsi su una superficie del diametro di parecchi chilometri. Poiché Norris e Dohl erano in grado di riconoscere alcuni individui, scoprirono che un determinato gruppetto di delfini aveva la tendenza a rimanere unito, ma spesso si spostava come una unità da un branco all'altro.

Secondo Norris e Dohl i delfini formerebbero gruppi strettamente uniti mentre stanno riposando allo scopo di impiegare le capacità sensoriali di tutti gli individui del branco per scandagliare l'ambiente circostante e scoprire un pericolo potenziale. È ben noto che i delfini riescono a scandagliare l'acqua per mezzo della ecolocalizzazione a una distanza molto maggiore di quella raggiungibile per mezzo della vista. (Nella ecolocalizzazione un delfino emette suoni di alta frequenza in brevi impulsi, in un modo simile a quello dei pipistrelli; i suoni rimbalzano sugli oggetti producendo echi che, captati di ritorno, forniscono informazioni sulla distanza, sulle dimensioni, sulla forma e persino sulla struttura dell'oggetto.) L'ipotesi di Norris e Dohl sulle capacità sensoriali combinate, particolarmente durante il riposo, si basa sull'abilità dei delfini di ricavare informazioni mediante ecolocalizzazione. È probabile che ogni singolo delfino all'interno di un branco strettamente organizzato possa captare i suoni di ecolocalizzazione emessi da altri membri del gruppo. Perciò, anche se ogni singolo individuo non produce molti suoni, le indicazioni sulla natura dell'ambiente circostante sarebbero captate con rapidità ed efficienza da tutti i membri del gruppo. È anche probabile che un gruppo in riposo nuoti nei pressi delle coste per rimanere in acque poco profonde, dove i grandi squali d'alto mare di solito non si spingono.

Durante i periodi di vigilanza, le distanze tra i delfini della specie *Stenella longirostris* aumentano e gli animali compiono molti salti. Questa attività può costituire solamente un gioco, come molti hanno affermato, ma può rappresentare anche una forma di comunicazione. Un salto è di solito seguito da un forte schiaffo o tonfo, prodotto dal corpo del delfino che rientra in acqua. Questi suoni percorrono distanze abbastanza notevoli sott'acqua e forse indicano la presenza di delfini che saltano ad altri delfini. In effetti talvolta i branchi di talune specie di delfini convergono su un branco di animali che saltano vivacemente da una distanza di parecchi chilometri. Anche durante il pasto notturno nelle acque di profondità i delfini compiono molti salti al di fuori dell'acqua e producono tonfi rumorosi. In questo periodo i membri del gruppo sono sparsi su una vasta area: l'ipotesi che ancora una volta si affaccia è che l'attività acrobatica sia utile ai delfini per comunicare la propria posizione e forse anche tutte le informazioni riguardanti il numero dei delfini nelle vicinanze e le attività di ognuno.

Graham Saayman e C. K. Tayler, del Museo di Port Elisabeth in Sudafrica, hanno studiato i tursiopi dell'Oceano Indiano *(Tursiops aduncus)*, mentre invece io e mia moglie Melany Würsig ci siamo occupati dei lagenorinchi scuri dell'Atlantico meridionale *(Lagenorhynchus obscurus)*. Le due specie possiedono habitat simili a quello di *Stenella longirostris* delle Hawaii: tale habitat si può definire costiero-pelagico, in quanto tutte e tre le popolazioni si possono spesso osservare e studiare dalla costa, ma possono migrare lontano dalla costa, di solito in cerca di cibo. Le tre popolazioni mostrano modelli di comportamento e di migrazione del

Nel diagramma sono posti a confronto il peso del cervello e il peso corporeo dell'uomo, del delfino e dello scimpanzé. Il cervello dell'uomo pesa circa il 2 per cento del suo peso corporeo, quello dello scimpanzé meno dell'1 per cento e quello del delfino leggermente di più dell'1 per cento. Benché confrontando il rapporto peso del cervello/peso del corpo spesso si possa ottenere un'idea grossolana del livello comparato dell'intelligenza, si devono tuttavia tenere presenti anche altri fattori come la lunghezza del corpo, il modo con cui vengono usati gli arti e la complessità del cervello. Non è quindi possibile essere sicuri, solo sulla base d'un semplice confronto dei rapporti peso del cervello/peso del corpo, che un determinato animale sia più intelligente d'un altro.

tutto simili: perciò si pensa che l'habitat condizioni notevolmente il modo di vivere di questi mammiferi marini.

Si può verificare questa affermazione studiando i delfini in diversi ambienti. Si possono prendere in esame tre ambienti che interessano i delfini: oceano profondo, zona costiera oceanica, acqua dolce. Le popolazioni studiate in modo più approfondito sono quelle costiere. Susan H. Shane della Texas A & M University ha osservato i tursiopi dell'Oceano Atlantico al largo del Texas; A. Blair Irvine, Randall S. Wells e Michael Scott dell'Università della Florida li hanno osservati al largo della Florida; Melany Würsig e io li abbiamo osservati al largo delle coste argentine; Saayman e Taylor hanno osservato invece i delfini dalla gobba che vivono nell'oceano Indiano e nel Pacifico (del genere *Sousa*) al largo del Sudafrica. Tutte queste ricerche possono essere riassunte dal nostro studio sui tursiopi delle coste dell'Argentina.

Il tursiope troncato, una specie di delfini che vive in un habitat costiero e ha un'area di distribuzione estesa praticamente su tutto il globo, può essere osservato agevolmente dalla riva del mare. Per un periodo di 21 mesi mia moglie e io abbiamo studiato un branco di tursiopi che frequentava il mare vicino alla riva (sempre entro una distanza di un chilometro), in acque profonde al massimo 40 metri. Le nostre osservazioni sono state compiute dalle scogliere affacciate al mare e da un piccolo canotto di gomma, mediante registrazione dei suoni compiuta sott'acqua, fotografando le pinne dorsali allo scopo di riconoscere i singoli individui e seguendo i movimenti degli animali con un teodolite dalla spiaggia. (Con un teodolite si può determinare esattamente la posizione, il movimento e la velocità degli animali selvatici: il metodo è stato perfezionato e sviluppato da Roger Payne della New York Zoological Society e si è dimostrato utilissimo negli studi sugli spostamenti costieri dei mammiferi marini.)

Il branco era costituito da un numero di delfini variabile tra 8 e 22, in media 15. Alcuni individui stavano costantemente assieme, altri erano assenti di tanto in tanto. Perciò la popolazione era costituita da subunità di comunità variabili.

Alla fine i nostri avvistamenti erano stati effettuati su tutti gli animali che noi avevamo identificato precedentemente.

L'illustrazione schematizza la gerarchia di dominanza d'un gruppo di delfini in cattività, studiato da Gregory Bateson dell'Oceanic Institute di Hawaii. Il gruppo era costituito da due delfini macchiati *(grigio scuro)* e da cinque esemplari di *Stenella longirostris (grigio chiaro)*. Nel disegno i maschi guardano verso sinistra, le femmine verso destra. Le frecce indicano i segnali di minaccia eseguiti da un animale contro l'altro. Il più grande, Kahili, e Limu non sono mai stati oggetto di minaccia. Akamai, il delfino maschio più piccolo al fondo della gerarchia, pur non avendo mai minacciato alcun altro membro del gruppo, riuscì ad accoppiarsi con la femmina Lei.

Abbiamo concluso che conoscevamo tutti o quasi i 60 delfini che in un momento o in un altro facevano parte del gruppo. Benché tutti gli animali nel gruppo più grande interagissero tra loro in qualche modo, non siamo mai riusciti a osservare il gruppo riunito. Perciò è probabile che altri sottogruppi, pure costituiti da individui a noi noti, si spostassero lungo altri tratti di costa. Evidentemente questi animali hanno una grande area di diffusione: una volta abbiamo avvistato un sottogruppo di animali a noi noti a più di 300 chilometri a sud della nostra base. È probabile che la grandezza di un sottogruppo rappresenti un compromesso equilibrato tra il numero di animali necessari per la difesa contro i predatori, il numero adatto per una alimentazione efficiente e il numero più adatto per una interazione sociale soddisfacente, per la riproduzione e per la sopravvivenza dei più giovani.

In autunno, inverno e primavera i tursiopi sembravano nutrirsi di banchi di acciughe, attorno alle ore del mezzogiorno, a una profondità tra i 15 e i 35 metri. In questo periodo avanzavano come un branco sparpagliato, con ogni individuo separato dal più vicino di circa 25 metri. Dopo parecchi minuti trascorsi in questa formazione cominciavano a immergersi in profondità e a girare in cerchi su una determinata superficie. Per analogia con il comportamento dei lagenorinchi scuri, che in questa regione si cibano esclusivamente di acciughe, abbiamo pensato che i tursiopi troncati spingessero i banchi di acciughe verso la superficie dell'oceano e lì se ne cibassero. La formazione sparsa prima del pasto probabilmente ha lo scopo di scandagliare acusticamente la superficie più grande possibile alla ricerca di acciughe.

Questo tipo di cooperazione tra i membri di un branco di delfini è completamente diverso dal comportamento di questi animali durante l'estate. Nelle acque costiere dell'Argentina in questa stagione non si trovano acciughe: perciò i delfini si cibano principalmente di pesci solitari e di grandi dimensioni, che vivono tra le rocce presso la costa. In questo periodo i delfini si spostano in acque profonde da due a sei metri e si sparpagliano su una linea più lunga che larga, con ogni animale praticamente alla medesima profondità e il più vicino possibile alla costa. I delfini fiutano il percorso da seguire tra le rocce e si infilano nelle spaccature in cerca di preda.

Mentre il numero medio di delfini per ogni gruppo era di 15, durante l'estate questa media diminuiva (14) e, in inverno, aumentava a 20. Se ne deduce che è necessario un numero minore di delfini per avere successo nella ricerca individuale del cibo presso la costa. In effetti, è possibile che la risorsa - ossia i grandi pesci che vivono tra le spaccature della costa - sia limitata e che il numero di 14 delfini presenti nel medesimo tempo in una determinata area rappresenti il numero massimo della ricettività dell'area. Il limite è invece superiore quando i delfi-

ni cercano il cibo in cooperazione nei banchi di pesci.

Si devono considerare anche altre ragioni di queste fluttuazioni stagionali nelle dimensioni dei sottogruppi. Per esempio, si potrebbe pensare che le dimensioni aumentino perché i delfini sono più soggetti agli attacchi delle orche quando vanno in cerca di cibo più lontano dalla costa e che con sottogruppi più numerosi si possano proteggere meglio. Forse le fluttuazioni riflettono la massima attività negli accoppiamenti e nella riproduzione. Tuttavia, sulla base dei dati attualmente disponibili, sembra che il cibo sia il fattore più importante che incide sulle dimensioni (e presumibilmente sulla composizione) dei sottogruppi di tursiopi.

Passando ora a considerare la struttura interna dei sottogruppi, il fatto che i singoli delfini si spostino qua e là da un sottogruppo a un altro fa pensare che questi animali siano caratterizzati da una società che i mammalogi definiscono aperta. Tra i mammiferi terrestri lo scimpanzé africano *(Pan troglodytes)* possiede un sistema sociale molto simile. Anche taluni ungulati africani possiedono un certo grado di apertura, con individui che si spostano frequentemente da un sottogruppo all'altro, nell'ambito di un branco più rigidamente definito. Tra gli scimpanzé la variazione sembra essere una risposta alla disponibilità di cibo; gli animali cercano il nutrimento in piccole unità ma si aggregano in unità di maggiori dimensioni quando il cibo è stato trovato. Melany Würsig e io abbiamo trovato una situazione simile tra i lagenorinchi scuri; tuttavia non è ancora chiaro fino a qual punto una distribuzione irregolare di cibo possa influire sulle variazioni nelle dimensioni numeriche dei branchi dei tursiopi.

Nonostante la tendenza all'apertura, sono stati quasi sempre osservati gruppi di cinque o sei delfini spostarsi assieme. Un gruppo di tal fatta comprendeva un animale notevolmente grosso e un secondo adulto che si spostavano con un piccolo durante tutto il periodo di osservazione di 21 mesi: ho perciò dedotto che il secondo adulto fosse una femmina. L'animale di grandi dimensioni era l'unico che sbattesse costantemente la coda contro la superficie dell'acqua quando si avvicinava una barca. Gli allevatori esperti di delfini dicono che lo sbattimento della coda può significare ira, aggressione e avvertimento. Sarei portato ad affermare che il grande delfino fosse in un certo senso il capo del gruppo, ma non possiedo ancora dati sufficienti sul sistema sociale e sul comportamento dei delfini per convalidare tale ipotesi.

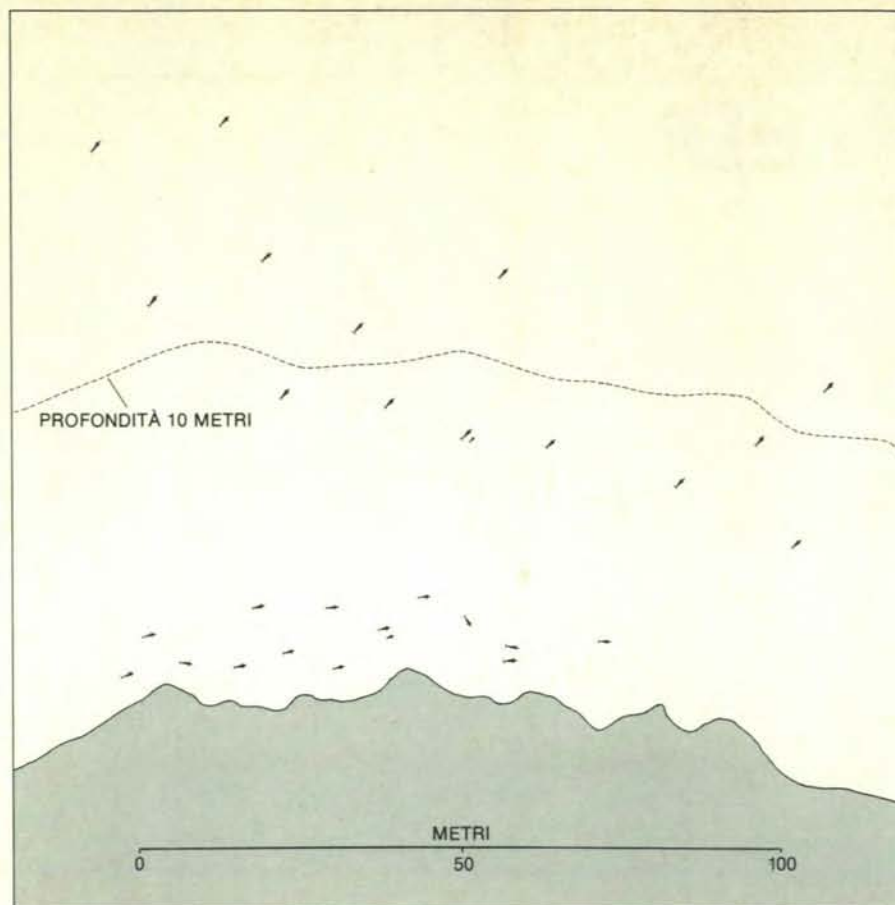
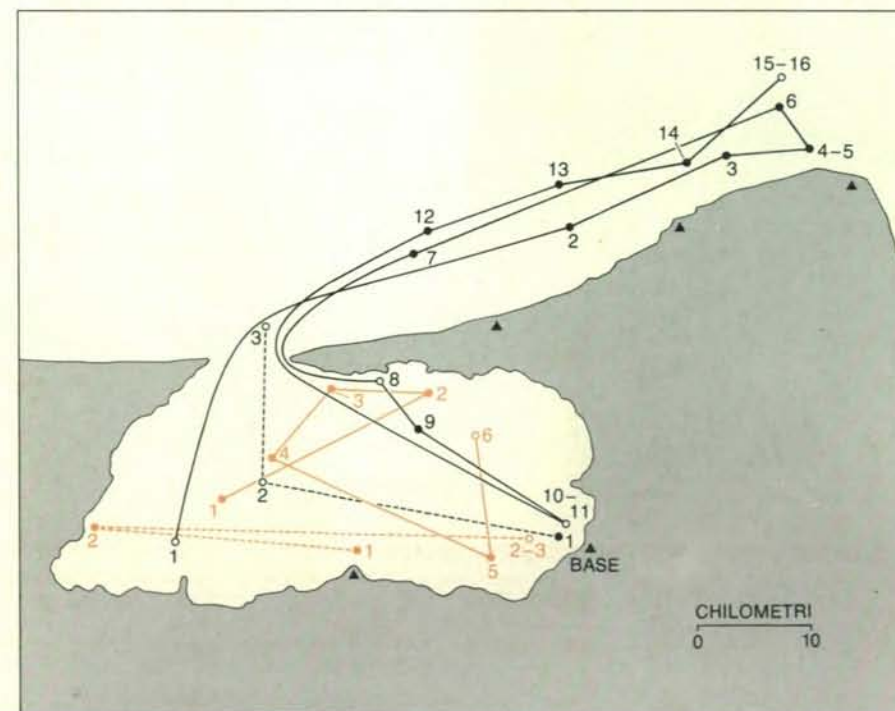
Due altri gruppi di cinque o sei delfini che abbiamo sempre trovato assieme erano costituiti da adulti approssimativamente delle stesse dimensioni. È possibile, a mio parere, che i gruppi fossero costituiti dai membri della popolazione non ancora in periodo riproduttivo, più o meno come i branchi di «elefanti scapoli» che si spostano assieme.

I dati più particolareggiati sulla composizione per classi di età e di sesso dei branchi di delfini ci vengono forniti dalla ricerca compiuta da Irvine, Wells e Scott sui tursiopi della Florida. Questi ricercatori catturarono 47 delfini e li contrassegnarono con targhette o altri sistemi, per identificarli in un secondo tempo nel loro habitat naturale. Durante le operazioni di contrassegno poterono determinare il sesso e le dimensioni di ognuno.

Dopo aver liberato i delfini, Irvine, Wells e Scott accertarono che l'area del branco residente era estesa su 85 chilometri quadrati. Le femmine e i piccoli spesso si spostavano in branchi che comprendevano solo alcuni maschi adulti, oppure nessun maschio adulto. Questi maschi tendevano ad associarsi più spesso con femmine senza piccoli che con madri e figli e, raramente, si riunivano a maschi subadulti. I maschi subadulti si trovano a volte in gruppi di scapoli, lontani dagli altri delfini. Furono osservate parecchie femmine con i loro piccoli per periodi di tempo lunghi, anche di 15 mesi: esiste quindi tra madre e figlio un forte legame sociale, che continua a lungo dopo lo svezzamento. Non è stata osservata invece una associazione di così lunga durata tra un maschio e un piccolo.

Si dovrebbe sottolineare che le relazioni sociali dei delfini non sono nette o immutabili: sono anzi notevolmente variabili. Tuttavia, sono evidenti alcune caratteristiche salienti: mentre i legami tra madre e piccolo sono forti, quelli tra maschio e femmina e tra maschio e piccolo non lo sono. Questo confronto fa pensare che l'accoppiamento sia in un certo senso promiscuo. I maschi subadulti possono essere esclusi dalla normale routine sociale, ma le femmine subadulte non lo sono; questo fatto indica che i maschi

I singoli delfini possono essere identificati dall'aspetto delle tacche e delle cicatrici del bordo posteriore della pinna dorsale. Queste fotografie sono un campione di immagini di 12 tursiopi troncati su un gruppo di circa 50. Tutti gli individui di questo gruppo poterono essere identificati.

La coesione di un gruppo di delfini è dimostrata da questa tabella su cui sono indicate le presenze e le assenze di 53 tursiopi troncati noti nelle acque costiere dell'Argentina, durante un periodo di 18 mesi. I termini a sinistra sono i nomi abbreviati dati ai delfini. La casella in colore a fianco del nome indica che quell'esemplare è stato visto presso la costa almeno una volta nel mese corrispondente; la casella bianca significa che l'animale non è stato visto. Un delfino indicato come avvistato può essere stato visto più di una volta durante il mese.

Le formazioni di un gruppo di tursiopi troncati lungo le coste dell'Argentina variavano a seconda che gli animali erano in acque poco profonde vicino alla costa o in acque profonde più al largo. Nelle acque poco profonde i delfini cacciavano individulmente i pesci di scoglio, mentre nelle acque profonde andavano a caccia in gruppo per scovare e raggruppare i banchi di acciughe.

Mediante radiorilevamento sono stati registrati gli spostamenti di quattro delfini al largo dell'Argentina (nel golfo di San José e in mare aperto). Ogni delfino è rappresentato da una traccia differente e il numero accanto a ogni pallino mostra la posizione dell'animale giorno dopo giorno, a partire dal momento in cui è stato equipaggiato con una radiotrasmittente. I pallini pieni rappresentano le posizioni note, ottenute sia per triangolazione dai punti della costa indicati da triangoli sia avvicinandosi ai delfini con una barca, mentre i pallini vuoti rappresentano le posizioni presunte. La posizione giornaliera degli animali è stata rilevata attorno a mezzogiorno.

adulti possono accoppiarsi maggiormente con varie femmine, rispetto a quanto possono fare le femmine con maschi diversi. Questo tipo di rapporti è caratteristico di un sistema di accoppiamenti poligino, che si verifica comunemente anche tra i mammiferi terrestri.

Circa il comportamento dei delfini i dati osservati sono molto meno numerosi rispetto a quelli sulla loro organizzazione sociale, benché sia stata raccolta una massa notevole di osservazioni sui delfini sia in natura che in cattività. Io credo che l'argomento non si conosca ancora abbastanza perché si possano enunciare osservazioni nette e generali. Ciò nondimeno, si possono citare alcuni esempi importanti di tipo di comportamento. I tursiopi troncati (e altre specie di delfini) s'impegnano nel corteggiamento e nell'accoppiamento lungo tutto il corso dell'anno, come è spesso indicato nei delfini in libertà dal fatto di nuotare pancia a pancia. Tuttavia nei tursiopi troncati e in alcune altre specie esiste un evidente massimo di nascite durante l'anno, e talvolta due massimi. Tra i delfini da noi osservati lungo le coste argentine tutti i piccoli nascevano in estate. Questo dato indica che esiste un cambiamento fisiologico nel maschio o nella femmina che fa sì che il concepimento avvenga in un periodo limitato. Questo cambiamento è stato documentato nell'aumento stagionale del peso dei testicoli nei maschi di parecchie specie di delfini.

Il fatto che gli accoppiamenti si susseguano durante tutto l'anno suggerisce anche che il corteggiamento potrebbe avere una connotazione diversa da quella sessuale. Parecchi ricercatori hanno suggerito che le interazioni sessuali possono anche avere la funzione di definire e di rafforzare le gerarchie e i legami sociali. L'affermazione è rafforzata dalla frequente attività omosessuale che si osserva almeno tra i delfini in cattività. Gli studi futuri forse dimostreranno che il termine più adatto dovrebbe essere «omosociale». L'utilizzazione di segnali sociali per determinare le gerarchie di dominanza è osservabile in molti altri gruppi di mammiferi.

Un secondo tipo di comportamento caratteristico di quasi tutte le specie di delfini è il salto. Ho già detto che questo comportamento di solito si verifica più spesso quando gli animali si trovano separati, cosicché potrebbe avere una funzione di comunicazione. I tursiopi troncati delle acque costiere dell'Argentina saltavano molto meno dei lagenorinchi scuri dello stesso ambiente, anche quando entrambe le specie cacciavano pesci più o meno allo stesso modo. Tuttavia i tursiopi si spostavano in un unico branco e perciò avrebbero avuto meno ragioni per comunicare con gli altri membri mediante salti.

I lagenorinchi si spostavano in circa 30 piccoli branchi in stretta vicinanza e abbiamo potuto spesso vedere tali branchi fondersi quando cominciavano i salti. I salti indubbiamente hanno altre funzioni, come quella di raggruppare o di catturare le prede, ma la funzione di comunicare tra

i membri di una particolare specie può essere importante.

Sulla base di una massa relativamente modesta di conoscenze, che cosa si può dire circa l'intelligenza dei delfini? Questi animali sono certamente abili nell'apprendere esercizi complicati, come dimostrano nei giochi di abilità negli acquari marini, e ricordano per parecchi anni gli esercizi appresi. Si è anche dimostrato che sono in grado di formulare pensieri relativamente astratti. Per esempio, Karen Pryor durante un lavoro svolto presso l'Oceanic Institute di Hawaii, è riuscita ad addestrare delfini dai denti rugosi *(Steno bredanensis)* a compiere un nuovo esercizio con la ricompensa di pesce. Dopo parecchi giorni di addestramento, i delfini dimostravano di saper compiere tipi sempre diversi di salti e contorsioni, evidentemente «rendendosi conto» che le forme di comportamento esibite precedentemente non sarebbero state ricompensate. Tuttavia, vari ammaestratori hanno fatto notare che la medesima cosa accade tra i cani e altri mammiferi e forse anche tra i colombi: ciò sta a indicare che non è necessario fare appello a una intelligenza di tipo superiore per spiegare ciò che i delfini sanno fare.

Edward O. Wilson della Harvard University ha suggerito che il cervello del delfino potrebbe essere più grande, rispetto alle dimensioni e al peso corporeo, del cervello della maggior parte degli altri mammiferi a causa della famosa capacità imitativa, che ha fatto dei delfini i beniamini degli ammaestratori e del pubblico. La domanda allora va così posta: come riesce questo animale a essere così abile nell'imitazione? R. J. Andrew dell'Università del Sussex ha notato che l'imitazione vocale può essere importante per quegli animali che spesso si spostano lontano l'uno dall'altro, in modo da uscire reciprocamente dalla visuale. Gli individui di un gruppo ampiamente diradato potrebbero quindi riconoscere gli altri membri grazie a una elaborata convergenza di segnali tra gli animali di un gruppo o di un branco. È stato dimostrato che questo sistema, sotto forma di dialetti, esiste presso alcuni primati e uccelli. È plausibile che il sistema si sia sviluppato per evoluzione fino a raggiungere una perfezione ancora maggiore tra i delfini, mammiferi che basano la loro comunicazione principalmente sui segnali acustici.

A proposito dell'imitazione dei movimenti, Wilson fa l'ipotesi che i singoli delfini potrebbero imitare quei membri del gruppo che si dimostrano più abili nel catturare i pesci e nell'evitare di cadere vittima dei predatori. Inoltre, è vantaggioso per gli animali sociali che cooperano per andare a caccia di cibo (come è stato dimostrato in alcune specie di delfini) conoscere bene i movimenti dei compagni di gruppo ed è altresì vantaggioso per i singoli animali assumere diversi ruoli nel raggruppare un banco di pesci. Wilson ne deduce che la sola capacità d'imitazione è sufficiente per spiegare le dimensioni del cervello del delfino e che i segnali sociali dei delfini non sono probabilmente più perfezionati di quelli della maggior parte dei mammiferi e degli uccelli. Secondo me non si sa ancora abbastanza sui segnali sociali dei delfini per poter dare un fondamento a questa osservazione. Norris considera l'abilità imitativa dei delfini non necessariamente migliore di quella di altri mammiferi con cervello più piccolo e fisicamente meno complesso. Ci sembra comunque futile attualmente confrontare l'intelligenza dei delfini con quella di altri mammiferi semplicemente per la mancanza di dati appropriati sui delfini e per l'enorme differenza tra l'ambiente di questi animali e quello dei mammiferi terrestri.

La necessità di acquisire migliori informazioni sui delfini induce alla ricerca dei metodi migliori per ottenerle. Una possibilità è quella di tentare di abituare i delfini alla presenza di osservatori, in modo che questi animali possano svolgere le attività giornaliere come se gli osservatori non ci fossero. George Schaller ha studiato in questo modo i gorilla di montagna e Jane Goodall ha aperto nuovi orizzonti sul mondo degli scimpanzé. Questi due ricercatori si spostavano con gli animali e stavano immobili pazientemente finché gli animali non li accettavano oppure semplicemente li ignoravano.

Ma come si fa a seguire un gruppo di delfini nell'oceano? Forse non è necessario. Jody Solow dell'Università della California a Santa Cruz recentemente ha imparato a fare un suono sott'acqua che talvolta richiama presso di lei gli individui d'un gruppo di delfini della specie *Stenella longirostris* delle Hawaii. Questa tecnica apre notevoli possibilità: un ricercatore potrebbe finalmente riconoscere tutti i membri di un gruppo, studiarne i modelli sociali e le interazioni e acquisire così una conoscenza più approfondita sul comportamento naturale dei delfini.

La gran varietà di salti che i delfini eseguono è esemplificata da queste fotografie di tre lagenorinchi scuri *(Lagenorhynchus obscurus)* in mare aperto. Si è potuto osservare che i delfini in cattività aumentano la varietà dei loro salti quando ricevono ricompense sotto forma di cibo.

Lo stetoscopio

Questo strumento, il primo d'uso generale in campo diagnostico, trasformò nel XIX secolo la pratica della medicina, modificando il quadro che il medico si faceva della malattia, e il rapporto tra paziente e medico

di Stanley Joel Reiser

Nel 1816 un giovane medico francese, René Théophile Hyacinthe Laënnec, fu chiamato al capezzale di una giovane donna che presentava i sintomi di una malattia cardiaca. A quell'epoca la maggior parte dei medici formulava la diagnosi osservando l'aspetto del paziente e interrogandolo sui sintomi. Trovandosi di fronte a una diagnosi insolitamente difficile, Laënnec fece ricorso a due metodi poco usati che coinvolgevono il contatto fisico con la paziente: la palpazione manuale del suo corpo per apprezzarne le strutture interne e la percussione del torace con le dita per generare suoni in grado di indicare le condizioni degli organi interni. Entrambe le operazioni furono ostacolate dall'obesità della paziente, mentre la morale del tempo proibiva a Laënnec di eseguire una auscultazione diretta, cioè di appoggiare l'orecchio sul torace per percepire i rumori del cuore. Imbarazzato, si ricordò che i suoni vengono amplificati se trasmessi attraverso determinati materiali. Egli arrotolò un foglio di carta in forma di cilindro e, collocandone una estremità sul torace della paziente, appoggiò l'orecchio all'altra estremità. Mediante questo strumento di fortuna egli percepì i rumori dell'attività cardiaca più distintamente di quanto non gli fosse riuscito fino ad allora.

Laënnec battezzò la sua invenzione con il nome di stetoscopio, dal greco *stethos*, «torace» e *skopein*, «vedere». Presto egli abbandonò il cilindro di carta in favore di quello che è stato poi chiamato lo stetoscopio di Laënnec: un cilindro di legno lungo circa trenta centimetri e con un diametro di circa quattro centimetri smontabile in due parti per essere facilmente trasportato. Una stretta fessura percorreva l'asse maggiore del cilindro: facendo esperimenti con prototipi diversi Laënnec aveva scoperto che una simile fessura esaltava le proprietà di trasmissione sonora dello strumento. La trasmissione di determinati suoni, come quelli dovuti alla respirazione, era ulteriormente migliorata da una apertura a imbuto all'estremità del cilindro collocata sul torace del paziente. Se dovevano essere indagati altri suoni, come quelli del cuore, un piccolo tappo di legno veniva inserito nell'apertura. Rendendosi conto che lo stetoscopio gli permetteva di percepire i suoni generati dal movimento di tutti gli organi della cavità toracica, Laënnec si mise all'opera per sviluppare una tecnica diagnostica sistematica basata sull'esame fisico condotto con lo strumento. Il risultato della sua precisa e convinta esposizione di questa tecnica, che egli definì auscultazione indiretta, fu che lo stetoscopio divenne il primo strumento a essere diffusamente usato dai medici nella diagnosi delle malattie.

La diagnosi, il processo attraverso il quale si determina la malattia di un paziente a partire dai suoi sintomi, è certamente un aspetto fondamentale della cura medica. La diffusione dello stetoscopio portò all'affermazione dell'esame fisico come chiave di volta della diagnosi. Esso operò una trasformazione profonda nella pratica medica, modificando sia la percezione della malattia da parte del medico sia il suo rapporto con il paziente. In particolare lo stetoscopio per la prima volta condusse il medico all'interno di una sfera privata in cui i segni gli erano comunicati direttamente dal corpo del malato. Esso permetteva al medico di fare una diagnosi obiettiva, ma lo allontanava anche dagli aspetti personali della malattia, aprendo una distanza che tutt'ora si va allargando. Certamente, oggi, lo stetoscopio stesso, che resta forse il più comune fra tutti gli strumenti medici moderni,è diventato vittima di quella stessa tendenza da esso inaugurata verso forme più obiettive di tecnologia diagnostica. Per capire come ciò sia avvenuto, è necessario prendere nuovamente in esame lo stato dell'indagine medica prima dell'invenzione dello stetoscopio.

All'inizio del XIX secolo, le diagnosi erano condotte soprattutto sulla base del resoconto verbale del paziente riguardo la sua malattia e sulle osservazioni da parte del medico (esclusivamente per mezzo dei suoi sensi) del colore della pelle del paziente, dello stato della sua lingua, della respirazione e così via. Nel XVIII secolo, tuttavia, alcune tecniche manuali - metodi di esplorazione diagnostica basati sull'esame fisico - avevano cominciato ad attirare l'attenzione. Una delle più significative, per quel che riguarda il modificarsi dell'atteggiamento verso la diagnosi clinica, era la tecnica della percussione che Laënnec aveva utilizzato con i suoi pazienti affetti da disturbi cardiaci: la percussione del corpo al fine di produrre suoni che avrebbero indicato il grado di vitalità degli organi interni.

La tecnica della percussione era un'invenzione del medico viennese Leopold Auenbrugger, che la descrisse per la prima volta nel suo *Inventum novum*, una piccola monografia pubblicata nel 1761. Auenbrugger aveva fatto notare, come Laënnec 58 anni più tardi, che le osservazioni dei segni esterni della malattia eseguite dal medico erano scarsamente utilizzabili nella diagnosi delle malattie toraciche. Pazienti affetti dalla stessa malattia mostravano spesso sintomi diversi, mentre pazienti affetti da malattie diverse esibivano spesso gli stessi sintomi. Auenbrugger era anche critico nei confronti del resoconto dei sintomi fatto dallo stesso paziente, da lui considerato mutevole e non degno di credibilità. Per Auenbrugger l'indice più sicuro della natura e del decorso della malattia toracica era il rumore patologico che riusciva a sollecitare mediante la percussione del paziente, e nell'*Inventum novum* egli affermò la propria capacità di formulare la diagnosi «in base alla testimonianza dei miei stessi sensi» (per tradurre dal suo latino).

Auenbrugger era interessato a trovare prove obiettive della malattia e riteneva che la tecnica della percussione, in quanto riduceva al minimo il ruolo della personalità e dell'aspetto fisico del paziente, avrebbe rivoluzionato la diagnosi delle malattie toraciche, offrendo ai medici un'alternativa sicura ai loro consueti metodi di diagnosi, poco affidabili. Sfortunatamente il metodo di Auenbrugger della deduzione della natura di una anomalia interna partendo dalle caratteristiche dei suoni sollecitati percuotendo il torace era in così stridente opposizione con le teorie correnti della medicina del tempo che durante la sua vita la percussione non ottenne mai l'apprezzamento che meritava.

Per adottare la percussione, un medico avrebbe dovuto pensare la malattia in termini anatomici, spiegando la genesi dei sintomi di un paziente in rapporto alle modificazioni strutturali dei tessuti del corpo. Questa rappresentazione della malattia era assai poco ortodossa in pieno XVIII secolo, quando la maggior parte dei medici abbracciava ancora la teoria umorale di origine greca, secondo cui la malattia era causata da uno squilibrio dei fluidi fondamentali del corpo, noti con il nome di umori. L'interesse per l'anatomia normale e le sue trasformazioni strutturali durante la malattia si era andato

Lo stetoscopio si è evoluto da strumento monoauricolare rigido fatto di legno a strumento biauricolare flessibile, fatto generalmente di gomma, plastica e metallo. Lo strumento in legno a sinistra è uno stetoscopio di Laënnec, che per primo nel 1816 usò un foglio di carta arrotolato per amplificare e convogliare alle proprie orecchie i suoni generati dal movimento degli organi all'interno del torace. Le proprietà di trasmissione del suono dello stetoscopio di Laënnec sono migliorate mediante una piccola fessura che corre lungo tutta la sua lunghezza e che si allarga all'estremità in una cavità a forma di imbuto che viene appoggiata al torace del paziente. Il piccolo pezzo di legno a sinistra in basso viene inserito nella cavità per migliorare la trasmissione di determinati suoni come quelli prodotti dal cuore. Uno stetoscopio biauricolare acusticamente superiore e anche flessibile cominciò a essere usato comunemente intorno al 1890. A destra, una versione moderna di questo strumento, lo stetoscopio o fonendoscopio di Littmann. L'estremità da appoggiare al torace porta su un lato una camera di metallo a forma di campana bordata di gomma per la trasmissione dei suoni di bassa frequenza e, dall'altro lato, una camera più svasata e appiattita coperta da un diaframma teso e sottile di plastica per la trasmissione di suoni di frequenza più elevata. Questo stetoscopio di legno si trova al Warren Anatomical Museum della Harvard Medical School; può darsi che sia appartenuto proprio a Laënnec.

evolvendo fin dal sedicesimo secolo, quando vennero pubblicati trattati rivoluzionari come il *De humani corporis fabrica* di Vesalio, ma i medici del XVII secolo pensavano ancora che le anomalie anatomiche fossero di importanza secondaria nel favorire la comprensione della genesi della malattia e della produzione dei sintomi. Un'opera che avrebbe profondamente influenzato i medici a pensare in termini di anatomia, il *De sedibus et causis morborum per anatomen indagatis* (Sedi e cause delle malattie scoperte mediante l'indagine anatomica) dell'anatomico italiano Giovanni Battista Morgagni, venne pubblicata nello stesso anno dell'*Inventum novum* di Auenbrugger ed ebbe quindi scarso effetto nei confronti delle prime reazioni al metodo della percussione. (All'epoca in cui Laënnec cominciò a proporre l'auscultazione indiretta, tuttavia, tentativi come quelli di Auenbrugger e di Morgagni si erano tradotti in un atteggiamento un poco più favorevole nei confronti della spiegazione della malattia in termini anatomici).

La mancanza di entusiasmo per il metodo percussivo era dovuta anche al modo con cui Auenbrugger lo presentava. Le variazioni dei suoni da percussione mediante le quali possono essere distinti i processi morbosi sono elusive. Il medico, per apprendere le modalità di applicazione del metodo, aveva bisogno di descrizioni esaurienti e di molti esempi relativi a questi complessi fenomeni acustici. Nel suo breve trattato, Auenbrugger preferì fornire soltanto delle spiegazioni brevi e poco precise, confidando che la sua scoperta sarebbe stata valutata «da coloro che sono in grado di apprezzare, in primo luogo, la scienza medica». Indubbiamente, egli si rifiutò di elaborare la sua tecnica per gli scettici e i critici, che egli giudicava essere «assaliti da invidia, malignità, denigrazione e calunnia».

La tecnica percussiva di Auenbrugger aveva diversi difetti. Per esempio, essa non era efficace in caso di pazienti affetti da malattie localizzate in profondità nella cavità toracica. Il confronto tra i suoni sani e di origine morbosa nel medesimo paziente costituiva una parte importante della diagnosi mediante percussione, per cui essa non funzionava bene nel caso di pazienti con entrambi i polmoni malati. I medici di oggi che usano la percussione appoggiano le dita di una mano sul corpo del paziente per attenuare il colpo ma la forma di percussione che Auenbrugger sosteneva era più violenta, e richiedeva che il medico portasse al corpo del paziente un colpo diretto con la punta di tutte le dita di una mano. Questo modo di procedere poteva risultare molto faticoso per il paziente, e alcuni medici si rifiutarono di adottarlo proprio per tale ragione. Inoltre, a quell'epoca, qualsiasi attività manuale era di solito considerata non confacente alla dignità del medico. In particolare l'intimità fisica richiesta dalla percussione minacciava di diminuire il livello professionale del medico, fino a collocarlo sullo stesso piano del chirurgo, sul quale egli affermava una superiorità sia tecnica, sia sociale.

Ai suoi tempi Laënnec si trovò di fronte proprio allo stesso genere di resistenze mediche e sociali, nei confronti dell'esame fisico manuale, che Auenbrugger aveva affrontato a metà del XVII secolo. C'era stato un cambiamento significativo dell'atteggiamento della medicina nel periodo di tempo compreso tra l'introduzione del metodo percussivo e l'introduzione dell'auscultazione indiretta. Nel 1816 i medici erano in genere disposti a credere che specifiche alterazione anatomiche costituissero segni sicuri di determinate malattie e che potessero essere usate per operarne la differenziazione. In particolare la correlazione tra i sintomi nel paziente vivo e le modificazioni strutturali trovate nel corpo nel corso dell'autopsia, stavano cominciando a svolgere un ruolo importante nella scoperta e nella verifica dei dati medici (e nell'addestramento dei medici). Quindi Laënnec poteva dimostrare mediante l'autopsia che se un suono di un certo tipo veniva percepito nel torace, lesioni di un certo tipo venivano trovate nel cadavere.

Come Auenbrugger, Laënnec riteneva che un complesso di segni sonori poteva costituire la base di una tecnica diagnostica nuova e più sicura. Egli trascorse tre anni all'Ospedale Necker di Parigi auscultando soggetti sani e malati, attento ai soffi, ai rantoli, ai sibili e ai gorgoglii all'interno dei loro toraci ed effettuando autopsie allo scopo di identificare e classificare i suoni. Nel 1819 pubblicò il *Traité de la auscultation médiate*, una critica di ampia portata dei metodi diagnostici dell'epoca che culminava in una descrizione esauriente della tecnica dell'auscultazione mediata. Il trattato è una esplorazione meticolosa dei rapporti tra i suoni generati all'interno del corpo, specifiche alterazioni anatomiche e malattie. L'opera di Laënnec, circa 10 volte più lunga della monografia di Auenbrugger sulla percussione, contiene descrizioni dei suoni che superano in chiarezza la maggior parte di quelle che sarebbero apparse in seguito. «Dopo che il paziente ha tossito o ha parlato, e ancor più durante la respirazione - scriveva Laënnec di un caso clinico - si udiva un tintinnio simile a quello di una piccola campana che ha appena cessato di suonare o di una zanzara che ronza dentro un vaso di porcellana». Nell'ottenere l'attenzione e la considerazione di una vasta platea di medici, l'auscultazione mediata doveva molto a tali vivide descrizioni.

Laënnec sviluppò anche una nuova tecnica sistematica per l'utilizzazione dello stetoscopio nella diagnosi delle malattie. Prima dell'invenzione dello stetoscopio le diagnosi si basavano principalmente sul resoconto che il paziente faceva dei propri sintomi e sulle osservazioni che il medico faceva del paziente; l'esame diretto del paziente aveva un ruolo scarso o nullo. La tecnica di Laënecc, da lui definita auscultazione mediata, forniva un approccio più sicuro alla valutazione della malattia perché si basava sull'analisi obiettiva dei segni acustici scoperti con lo stetoscopio.

All'inizio i medici trovarono una quantità di motivi per sottoporre a critica la tecnica. Alcuni pensavano di apparire ridicoli, piegati sopra il corpo dei loro pazienti, ascoltando suoni mediante uno strumento che ricordava una bacchetta divinatoria. E resisteva ancora la mentalità secondo cui l'uso di strumenti e di tecniche manuali era al di sotto dello status del medico. Il trasporto degli strumenti medici era particolarmente scomodo, e alcuni medici escogitarono degli espedienti, come portare le due parti smontate dello stetoscopio incrociate dentro il loro copricapo.(Questa usanza particolare non era senza rischi: in un caso, quando una palla di neve scoperchiò il copricapo di uno studente in medicina mostrando uno stetoscopio, lo studente fu accusato di portare un'arma nascosta.)

Inoltre, molti medici trovarono che le ipotesi dei primi sostenitori dello stetoscopio erano eccessive. Nel *Traité de la auscultation médiate* Laënnec giudicava i metodi dell'epoca - comprese le tecniche manuali - in modo aspro, trovando che la palpazione era «troppo vaga e incerta per essere di qualche aiuto» e la percussione appena un poco più valida. Egli si mostrò anche critico nei confronti della antica prassi di misurare i battiti del polso, giustificando la sua popolarità con il fatto che tale prassi non recava disturbo né al medico, né al paziente, e che qualsiasi irregolarità era soggetta a un'ampia gamma di interpretazioni. Laënnec e molti dei suoi seguaci affermavano che una diagnosi autorevole poteva essere condotta soprattutto - di fatto, a volte, esclusivamente - sulla base di informazioni auditive. Altri medici meno partigiani ritenevano che si potesse giungere a una diagnosi più affidabile basandosi parimenti sui sintomi auditivi e sulle informazioni ottenute mediante i metodi tradizionali: essi erano comprensibilmente seccati dalle ipotesi unilaterali di Laënnec e dei suoi seguaci. Di fatto le pretese eccessive legate allo stetoscopio erano in parte dovute alla straordinaria abilità di Laënnec. Possedendo un udito notevolmente acuto e una conoscenza particolareggiata del significato patologico delle più tenui variazioni di sonorità, egli era in grado di distinguere le malattie sulla base dell'informazione auditiva con straordinaria accuratezza. Pochi tra i seguaci di Laënnec riuscivano a valutare quanto poco comuni fossero le sue capacità.

Le reazioni allo stetoscopio da parte dei pazienti furono in genere favorevoli, sebbene alcuni fossero spaventati dallo strumento (poiché associavano l'idea di strumento con quella di chirurgo) o imbarazzati dalla intimità fisica richiesta per l'auscultazione. Alcuni rifiutarono di essere esaminati con lo stetoscopio a causa della pressione eccessiva con cui il loro medico premeva lo strumento contro di loro a causa dell'inconveniente di dover effettuare frequenti cambiamenti di posizione durante un tale esame. Ciò che i pazienti temevano maggiormente, tuttavia, era la precisione dello stetoscopio. Spesso esso trasformava l'ultima speranza che la malattia fosse guaribile nella certezza che non lo era.

In ultima analisi, ovviamente, la grande risorsa della auscultazione mediata era proprio la sua precisione e la sua affidabilità: la tecnica poteva convertire un sospetto diagnostico in una certezza. Come facevano notare Laënnec e i suoi seguaci, avere fiducia in una diagnosi significava che la malattia poteva essere trattata con rinnovato vigore. La terapia poteva essere adottata immediatamente, a volte nei primi stadi della malattia. E anche se la malattia diagnosticata era incurabile, si poteva perlomeno risparmiare al paziente una terapia inutile e forse dolorosa.

La precisione della diagnosi formulata con lo stetoscopio si tradusse in una crescente accettazione, da parte del pubblico, della auscultazione mediata. Di fatto, nell'arco di un decennio dalla pubblicazione del *Traité* di Laënnec, il sentimento popolare in favore di questa tecnica era diventato fortissimo e un medico che non avesse utilizzato lo stetoscopio metteva a repentaglio la propria reputazione professionale. Alcuni medici che non sapevano come auscultare i pazienti cominciarono a portare con sè lo stetoscopio semplicemente per esibirlo. L'esame fisico - attraverso l'auscultazione - costituiva il nuovo modo di far diagnosi.

Nella seconda metà del XIX secolo furono apportati diversi miglioramenti allo stetoscopio rigido e manoauricolare di Laënnec. Per accontentare i medici che si seccavano di dover trasportare un oggetto così voluminoso, furono progettati stetoscopi più piccoli. Modifiche apportate alla struttura dello strumento (per esempio, la creazione di una cavità parabolica al posto di quella conica all'estremità da appoggiare al torace del paziente) migliorarono le sue proprietà acustiche. Già nel 1829 il primo stetosco-

Lo stetoscopio parzialmente flessibile fu progettato nel 1829 da Nicholas Comins, un medico di Edimburgo. Fatto di due tubi rigidi *(a, b)* collegati da un giunto e uniti a un auricolare mobile *(e)*, lo strumento poteva essere piegato ad angolo a piacimento, così che il paziente non doveva modificare la propria posizione troppo spesso, durante una visita e, a sua volta, il medico non doveva piegarsi troppo. Furono infine sviluppati stetoscopi completamente flessibili; questi risparmiavano al paziente il disagio di avere uno strumento rigido premuto contro il corpo e diminuivano anche la necessità di uno stretto contatto fisico tra il medico e il paziente.

I rumori patologici del cuore sono suoni auscultatori confusi del cuore causati da difetti strutturali del tessuto cardiaco. In basso sono indicate intensità e durata di diversi suoni comuni di questo tipo (come sono sentiti in vari punti del torace); in alto è mostrato il flusso del sangue attraverso il cuore. Il sangue è pompato attraverso il corpo dalla contrazione dei ventricoli cardiaci. Il periodo durante il quale i ventricoli destro e sinistro sono contratti e il sangue viene forzato nell'arteria polmonare e nell'aorta è la sistole; il periodo durante il quale i ventricoli sono rilasciati è la diastole. In un singolo ciclo cardiaco di un cuore normale si distinguono due toni: dapprima i battiti di chiusura contemporanei delle valvole tricuspide e mitrale, che impediscono al sangue di rifluire negli atri durante la contrazione dei ventricoli; e poi i battiti non del tutto simultanei delle valvole polmonari e aortiche, che impediscono al sangue di refluire nei ventricoli quando sono rilasciati. Nel diagramma il rettangolo nero rappresenta il primo tono di ogni ciclo; il rettangolo grigio (la componente polmonare) e il rettangolo in colore chiaro (la componente aortica) rappresentano il secondo tono. I suoni di frequenza inferiore sono rappresentati da linee più spaziate. Se la valvola aortica è ristretta o ostruita (difetto noto come stenosi aortica) allora durante la sistole, quando il sangue è forzato attraverso la valvola, si genera una quantità abnorme di turbolenza e si produce un suono che si alza e si abbassa chiamato soffio da eiezione. Nell'insufficienza aortica non funziona la valvola aortica, così che durante la diastole il sangue refluisce nei ventricoli, producendo un suono abbastanza fisso chiamato soffio da rigurgito. Se c'è un difetto del setto atriale nel cuore (un foro nella parete che separa gli atri), allora durante la sistole una grande quantità di sangue passa attraverso di esso, facendo sì che le componenti aortica e polmonare del secondo suono cardiaco risultino ampiamente distanziate; durante la sistole il ventricolo destro espelle il sangue nell'arteria polmonare, provocando un soffio da eiezione. Un tipo diverso di suono viene generato quando la valvola mitrale è ristretta o ostruita (difetto noto come stenosi mitralica). In questo caso si produce uno schiocco di apertura della valvola mitrale *(rettangolo in colore più scuro)* all'inizio della diastole, seguito da un soffio diastolico da eiezione a bassa frequenza nel momento in cui il sangue fluisce nel ventricolo sinistro. I diagrammi sono tratti dal volume *Bedside Diagnostic Examination* di Elmer L. DeGowin e Richard L. DeGowin.

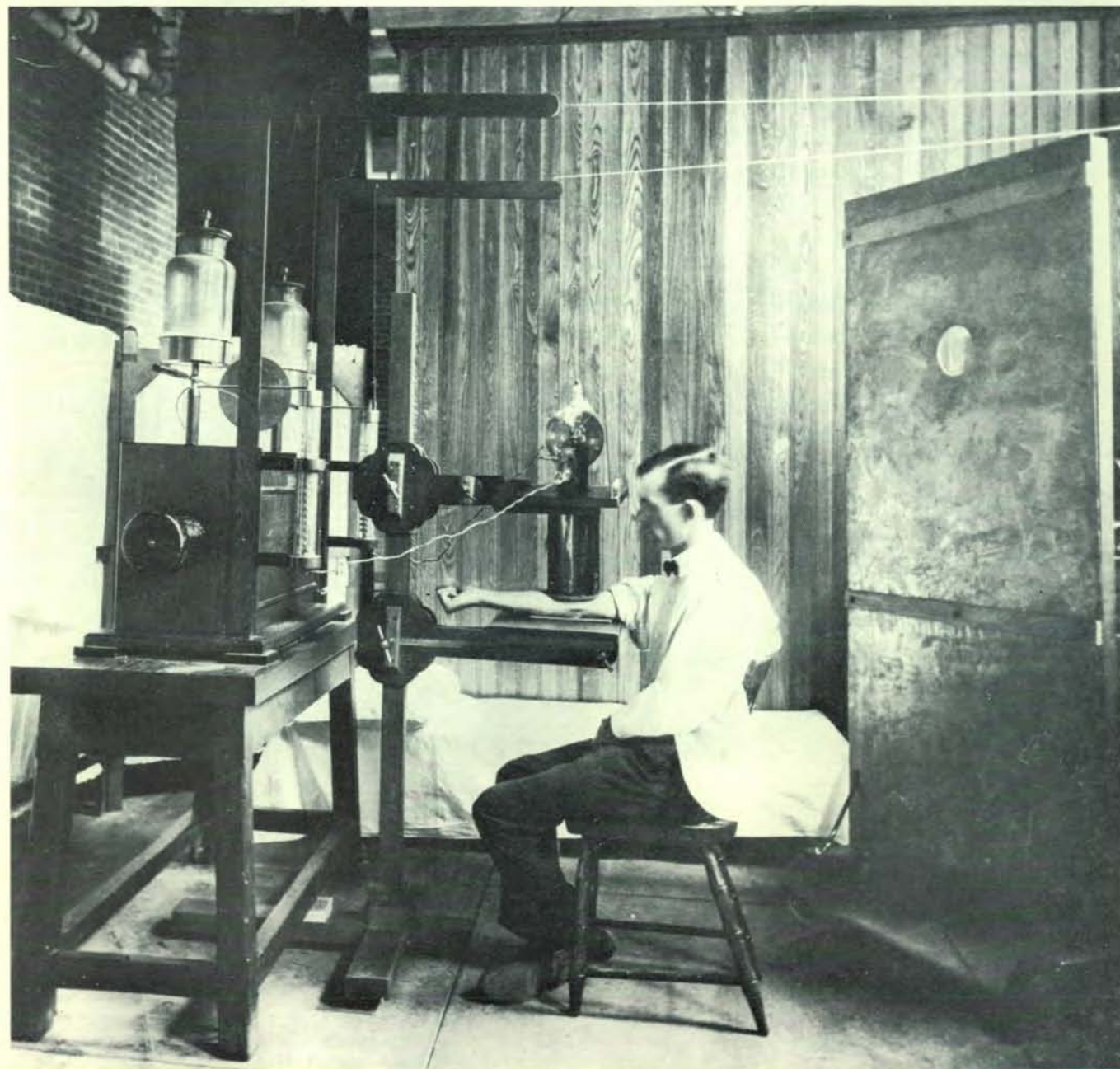
pio non completamente rigido venne costruito da un medico di Edimburgo, Nicholas Comins. Lo strumento era formato da due tubi di legno rigidi, ciascuno dei quali era lungo circa diciotto centimetri, uniti mediante un giunto articolare in modo da poter essere disposti secondo un angolo qualsiasi: un auricolare collocato a uno dei due tubi ruotava attorno a un giunto.

I primi stetoscopi manoauricolari completamente flessibili fatti di tubi pieghevoli comparve attorno al 1830. Questi strumenti, che potevano essere lunghi fino a settanta centimetri circa, evitavano ai pazienti di dover cambiare ripetutamente posizione durante l'esame e il disagio di uno strumento rigido premuto contro di loro. Inoltre, con uno stetoscopio flessibile, i pazienti non erano più costretti a un rapporto così ravvicinato con il medico: una donna poteva appoggiare da sè al proprio corpo la camera dello stetoscopio flessibile tenendo quindi il medico alla giusta distanza. Analogamente i medici potevano evitare un contatto ravvicinato con pazienti affetti da malattie contagiose. Inoltre potevano evitare il fastidio di curvarsi per esaminare i pazienti (un sollievo in particolare per i medici grassi) e potevano auscultare sè stessi.

Comins fu probabilmente anche il primo a proporre lo stetoscopio biauricolare: uno stetoscopio flessibile con i tubi che arrivavano a entrambe le orecchie. Tuttavia, verso la metà del secolo scorso furono introdotti i primi modelli funzionanti in modo soddisfacente. Nel 1878, quando fu inventato il microfono, sembrò possibile una amplificazione ancora maggiore dei suoni dell'organismo, ma divenne presto chiaro che i suoni cardiaci e polmonari captati dal microfono non erano abbastanza distinti per un uso diagnostico.

L'entusiasmo iniziale e la successiva delusione nei confronti del microfono coincisero con la introduzione di un gran numero di modelli diversi di stetoscopio, con le conseguenze di una grande confusione. I medici cominciarono a lamentarsi che qualsiasi amplificazione avrebbe alterato i suoni ai quali si erano abituati. Molti preferirono ignorare la grande disponibilità di stetoscopi differenti, insistendo sul fatto che era la scienza del medico che importava nell'auscultazione, e non lo strumento che egli maneggiava. Il risultato fu che lo stetoscopio biauricolare flessibile, acusticamente superiore (e comodamente trasportabile), non venne comunemente adottato fino agli anni novanta. Ai primi del 1900 i medici scoprirono che un diaframma teso sull'apertura della camera dello stetoscopio avrebbe migliorato la trasmissione di determinati tipi di suoni, per cui tale dispositivo fu incorporato allo strumento. Gli stetoscopi moderni o fonendoscopi seguono lo stesso progetto di base e in genere possiedono due camere: una a forma di campana, senza diaframma, per la trasmissione di suoni di frequenza minore e una più svasata, con diaframma, per la trasmissione di suoni di frequenza più elevata.

Paradossalmente, l'accettazione dello stetoscopio biauricolare coincise con la scoperta di una tecnica che avrebbe rapidamente soppiantato lo stetoscopio come fondamentale mezzo diagnostico nelle malattie toraciche. Nel 1895 W. K. Roentgen scoprì che un tubo a raggi catodici emetteva radiazioni in grado di attraversare un oggetto solido e di fissare l'immagine della struttura interna dell'oggetto su una lastra fotografica. La notizia della scoperta di Roengten di questi raggi X, un «nuovo genere di luce che attraversa e fotografa tutto», si diffuse rapidamente. L'immagine radiografica da lui fatta della mano della moglie fu riprodotta su pubblicazioni in tutto il mondo, e i giornali fecero a gara nel riportare resoconti esagerati sugli usi potenziali della scoperta. Il pubblico acquistò con avidità immagini radiografiche delle mani delle persone care e di oggetti d'uso quotidiano, come chiavi tenute in un contenitore opaco quale un borsellino. (La regina del Portogallo fece fare immagini radiografiche di tutti i membri della sua corte; le coste storte evidenziate dai raggi X indussero alcuni membri femminili del seguito a buttare via i loro corsetti strettamente allacciati.)

I medici erano straordinariamente entusiasti della scoperta di Roentgen e applicarono ben presto i raggi X alla diagnosi di fratture ossee e di calcoli renali e alla localizzazione di corpi estranei come pallottole infisse nei tessuti del corpo. Entro un mese dalla scoperta era in funzione un apparecchio che proiettava i raggi X attraverso il corpo di un paziente su uno schermo fluorescente, così che il medico poteva anche osservare direttamente il movimento del cuore e dei polmoni. Con il miglioramento della tecnica dei raggi X divenne possibile analizzare malattie a carico dei tessuti molli del-

In questa fotografia, scattata attorno al 1910 al Massachussetts General Hospital di Boston, compare una delle prime apparecchiature a raggi X. I raggi X furono scoperti nel 1895 (più o meno nello stesso periodo in cui si diffuse ampiamente lo stetoscopio biauricolare flessibile) e furono rapidamente utilizzati per una quantità di applicazioni mediche, come la diagnosi di fratture ossee e la localizzazione di calcoli renali. In particolare le malattie cardiache e polmonari potevano essere diagnosticate più precocemente e più accuratamente con i raggi X che non con uno stetoscopio. Il risultato fu che la macchina a raggi X cominciò a superare lo stetoscopio come strumento fondamentale nell'analisi delle malattie toraciche. Nel ventesimo secolo lo stetoscopio è stato sempre più trascurato, sia nella pratica sia nell'addestramento dei medici.

l'organismo come i polmoni e il cuore.

Con i raggi X i medici erano spesso in grado di individuare malattie a carico dei polmoni e del cuore più precocemente, e di definirne le caratteristiche con maggior precisione di quanto non potessero con lo stetoscopio. Nel 1896 il pioniere della radiologia, Francis H. Williams, espresse i sentimenti di molti altri medici con queste parole: «Possiamo ora guardare là dove prima potevamo soltanto ascoltare e qualche volta sentire, e solo in modo imperfetto». Per molti la trasformazione dei suoni percepiti nell'auscultazione in immagini nella mente del medico sembrava meno affidabile dell'analisi diretta delle alterazioni evidenziabili in un'immagine radiografica. Anche i pazienti riponevano maggiore fiducia in una diagnosi fondata sull'alterazione anatomica visibile in una immagine. L'uso dello stetoscopio cominciò a declinare, e la tendenza fu accelerata nel XX secolo dallo sviluppo di strumenti più sofisticati di tecnologia diagnostica come l'elettrocardiografo. Certamente i medici, rimanendo sempre più sedotti dalla tecnologia diagnostica, diventavano sempre meno interessati all'esame fisico in generale e alle tecniche sviluppate nel XIX secolo per eseguirlo.

Oggi lo stetoscopio è il vecchio cavallo di battaglia della medicina. Sebbene non possa competere con la schiera di tecnologie elaborate e costose alle quali ha spianato la strada, resiste tenacemente al pensionamento. La sua forza attuale, nei nostri tempi, si fonda in parte nel dare sia ai medici che ai pazienti il senso della continuità con il passato. Identificato con l'affidabilità della diagnosi, l'oggetto familiare evoca confidenza. Fatto più importante, esso fornisce a quei medici che ancora sanno come usarlo, informazioni attendibili, immediate e a basso costo, che possono eliminare la necessità di complicati test diagnostici.

È facile, naturalmente, per il medico moderno, inviare il proprio paziente a un laboratorio clinico o a un laboratorio radiografico per esami diagnostici, invece di sottoporlo a un ampio esame mediante lo stetoscopio. Tali esami di laboratorio sono in genere rimborsati dalle organizzazioni mutualistiche, per cui il paziente deve sopportare un costo immediato basso e il medico risparmia tempo. Certo, dal momento che pochi medici sono addestrati in metodologie come l'auscultazione, la scelta della strada della tecnologia è solo prudente. Ciò che viene perduta è, in primo luogo, la fiducia che il medico ha in se stesso. Quando il medico del XIX secolo scelse di effettuare la diagnosi basandosi meno sul resoconto verbale del paziente relativo ai sintomi, e più sui segni fisici della malattia che in molti casi solo lui era in grado di scoprire, era costretto a concentrare la propria mente sulla malattia. Il XX secolo dal punto di vista medico può essere caratterizzato come un'epoca in cui i medici, per raccogliere e valutare le prove della malattia, sono portati ad avere meno fiducia in se stessi e più negli specialisti, nei tecnici e nelle macchine.

GIOCHI MATEMATICI

di Martin Gardner

I piaceri delle circonferenze tangenti

Esistono centinaia di splendidi problemi a proposito delle circonferenze tangenti, molti dei quali fanno parte della matematica ricreativa. Alcuni sono già stati trattati in questa rubrica. Questo mese ne presenterò di nuovi. Non fornirò dimostrazioni per mancanza di spazio, ma se i lettori si sentiranno sfidati a trovarle non potranno fare a meno di affinare in questo modo la loro comprensione della geometria piana elementare.

Iniziamo con una figura famosa che si dice sia stata studiata per la prima volta da Archimede e nota come «arbelo», nome greco che indica il trincetto o coltello da calzolaio, dato che assomiglia alla forma della lama usata dagli antichi calzolai. Si tratta della regione sfumata nella figura in alto della pagina a fronte, delimitata dalle semicirconferenze di diametri AB, BC e AC. B può essere un punto qualsiasi su AC. Ecco alcune delle proprietà più interessanti di questa figura:

1. La lunghezza dell'arco semicircolare AC è uguale, come si dimostra facilmente, alla somma degli archi AB e BC.

2. Si tracci BD perpendicolare ad AC. L'area dell'arbelo è uguale all'area di una circonferenza di diametro BD.

3. BD divide l'arbelo in due parti. Le circonferenze inscritte in ognuna delle parti sono identiche dato che ogni circonferenza ha diametro $(AB \times BC) / AC$. La minima circonferenza che circoscrive queste due circonferenze gemelle ha area uguale a quella dell'arbelo.

4. Si tracci una retta tangente agli archi AB e BC. I punti di tangenza E ed F giacciono sulle rette AD e CD.

5. EF e BD sono uguali e si bisecano reciprocamente. Ciò assicura che la circonferenza di diametro BD passi attraverso i punti E ed F.

6. Agli inizi degli anni cinquanta, Leon Bankoff fece una curiosa scoperta. (Bankoff è un dentista di Los Angeles che dice di essere «di estrazione russa». Egli cura la rubrica dei problemi del «The Pi Mu Epsilon Journal» ed è noto per la sua abilità nell'inventare e nel risolvere problemi.) Bankoff ha precisato le nostre conoscenze a proposito delle circonferenze gemelle di Archimede mostrate da entrambi i lati di BD nella figura in alto della pagina a fronte: si tracci una circonferenza (tratteggiata nella figura) tangente alle tre circonferenze più grandi e quindi se ne tracci una più piccola (in colore) che passi per B e per i punti in cui la circonferenza tratteggiata tocca gli archi AB e BC. Anche questa circonferenza è identica a quelle gemelle di Archimede. La dimostrazione è nell'articolo di Bankoff *Are the Twin Circles of Archimedes Really Twins?* («Mathematics Magazine», vol. 47, n. 4, settembre, 1974) No, non sono realmente gemelle, risponde Bankoff; sono due circonferenze in un insieme di tre.

7. Si costruisca all'interno dell'arbelo una serie di circonferenze tangenti. Quella tratteggiata mostrata nell'illustrazione in alto della pagina a fronte è la prima; le altre si possono costruire sulla sinistra, tante quante se ne vuole, procedendo nel modo mostrato nella figura in basso. Numeriamo le circonferenze: C_1, C_2, C_3 e così via. I centri di tutte queste circonferenze giacciono su un ellisse. Il diametro di una di esse, per esempio C_n, è un n-esimo della distanza perpendicolare che separa il centro della circonferenza dalla linea di base ABC. Questo notevole risultato si trova nell'opera di Pappo di Alessandria, matematico del IV secolo, che lo presenta come un vecchio teorema.

La dimostrazione di Pappo è semplice se si fa uso della geometria proiettiva, invertendo l'intera figura considerando A come il centro di inversione. Questa trasforma le semicirconferenze AB e AC in rette parallele, e la serie diventa un insieme di circonferenze uguali limitate dalle due rette. Troverete una buona spiegazione di questo risultato nell'articolo *A Chain of Circles* di Rodney T. Hood («The Mathematics Teacher», vol. 54 n. 3 marzo, 1961) e una spiegazione più breve in *Topics in Recreational Mathematics* di J. H. Cadwell (Cambridge University Press, 1966). Pappo non conosceva i metodi della geometria proiettiva (che si sviluppò solo nell'ottocento) e quindi la sua dimostrazione è più complicata

8. Se AB è uguale a 2 e AC è uguale a 3, allora la serie delle circonferenze riserva altre sorprese. I diametri delle circonferenze della serie sono frazioni razionali uguali a $6/(n^2+6)$. Quindi C_1 è uguale a 6/7, C_2 a 3/5 e così via. Come mi ha indicato recentemente Norman Pos, il centro di C_2 giace sul diametro della circonferenza esterna che è perpendicolare ad AC. Inoltre i centri di C_2 e C_3 giacciono su una retta parallela ad AC, e ciò vale anche per i centri di C_1 e di C_6. Quest'ultimo risultato è un caso particolare di un teorema più generale. Se AB e AC sono numeri interi, e AC è uguale ad $AB+1$, i centri di ogni coppia di circonferenze i cui indici hanno un prodotto uguale ad $AB \times AC$, giacciono su una retta parallela ad AC. Quindi se AB è uguale a 3 e AC a 4, le coppie di circonferenze di indici 1 e 12, 2 e 6, 3 e 4, hanno i centri su una retta parallela ad AC. (Per una dimostrazione si veda l'articolo *On a Generalization of the Arbelos* su «The American Mathematical Monthly» di M. G. Gaba, vol. 47, gennaio, 1940.)

9. Se B divide AC secondo il rapporto aureo, si ottengono molte altre sorprendenti proprietà, che sono discusse nell'articolo di Bankoff intitolato *The Golden Arbelos* («Scripta Mathematica», vol. 21 n. 1, marzo, 1955). Dedicato agli editori: Bankoff ha un manoscritto non pubblicato di dieci capitoli riguardante l'arbelo, scritto in collaborazione col matematico francese Victor Thébault.

Strettamente connesso all'arbelo è un sorprendente teorema scoperto da Jacob Steiner, un matematico svizzero dell'ottocento, e illustrato nella figura in alto di pagina 124. Si tracci una circonferenza e al suo interno un'altra più piccola. Nello spazio tra le due si inscriva una serie di circonferenze. Nella maggior parte dei casi le circonferenze non formeranno esattamente un anello di circonferenze tangenti, dato che le circonferenze terminali si sovrapporranno. In alcuni casi, tuttavia, si formerà un anello perfetto, come quello mostrato nella figura dalle circonferenze a linee piene. Quando ciò si verifica l'anello viene detto catena di Steiner. Il teorema di Steiner stabilisce che se le due circonferenze date permettono la formazione di una catena di Steiner, ne permettono infinite. Ossia, comunque si tracci la prima circonferenza della catena, si darà inizio a una catena perfetta. Una seconda catena qualsiasi è mostrata tratteggiata nella figura.

Come prima, la dimostrazione più semplice richiede la geometria proiettiva. Basta effettuare un'inversione che trasformi le due circonferenze iniziali in due circonferenze concentriche. La catena di Steiner diventa allora una catena di circonferenze identiche che riempie la regione tra le circonferenze concentriche. Il libro di Cadwell, citato in precedenza, fornisce i dettagli.

Solomon W. Golomb, i cui contributi nel campo della matematica ricreativa compaiono spesso in questa rubrica, durante un viaggio compiuto lo scorso anno in Europa scoprì di avere in tasca una grande varietà di monete di dimensioni diverse. Gli si presentò quindi il seguente problema. Supponiamo che n monete di varie dimensioni formino una catena che circonda esattamente una moneta centrale, come è mostrato nell'illustrazione in basso a pagina 124. Se si cambia l'ordine delle monete nella catena, queste formeranno ancora una catena perfetta?

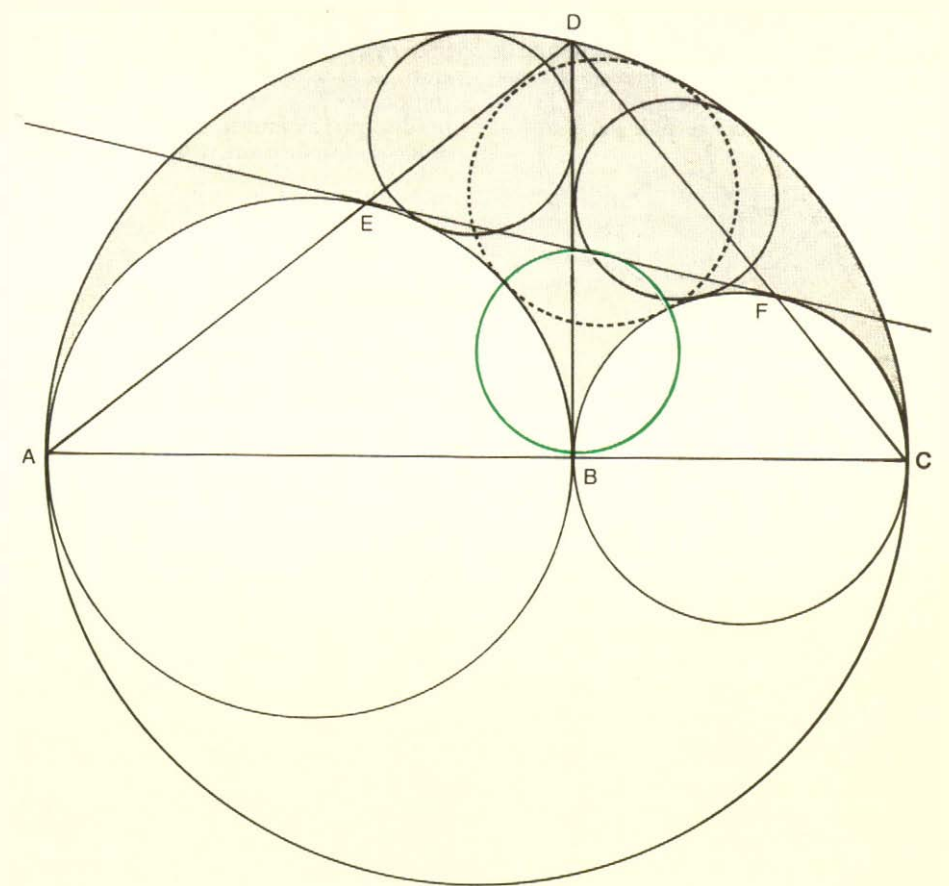
La maggior parte delle persone propende a rispondere sì, e vi è anche una «dimostrazione». Dal centro della moneta all'interno si traccino rette passanti attraverso ogni coppia di monete adiacenti, come mostrato nella figura. La somma di tutti questi angoli deve dare 360° e ciò sembra essere indipendente da come si dispongono le monete.

La dimostrazione è sbagliata, perché si suppone che i raggi siano tangenti a ogni coppia di circonferenze che toccano. Tuttavia non è sempre così, e in questi casi l'ordine in cui sono disposte le circonferenze può essere essenziale. Se sono date solo 3 monete, ovviamente un cambiamento nell'ordine non ha effetti dato che produce solo rotazioni o riflessioni della figura originale. Quando vi sono almeno quattro monete è facile trovare esempi in cui la catena chiude solo con certi ordinamenti e non con altri.

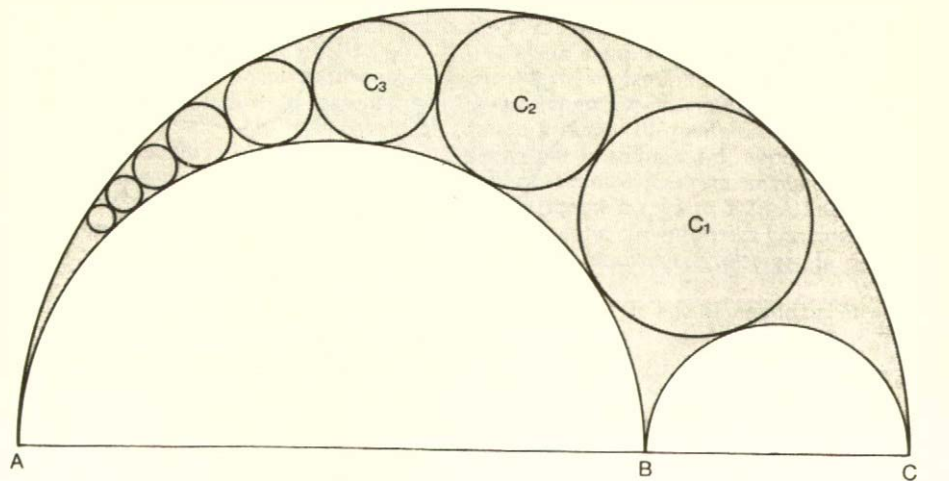
Le differenze che sorgono con le varie permutazioni sono minime, a meno che non vi sia una grossa differenza di dimensioni tra le varie monete. Quindi se disponete un mezzo dollaro, un quarto di dollaro, un nichelino, un decimo di dollaro e un penny intorno a una circonferenza data, troverete che ogni permutazione sembra dare luogo a una catena perfetta. Tuttavia vi sono differenze. Lascio ai lettori la dimostrazione di quest'ultima asserzione. Ecco un'altra delle scoperte di Golomb: date n monete tutte diverse, qual è il numero massimo di circonferenze diverse che possono racchiudere entro una catena cambiando il loro ordine? La soluzione è $(n-1)!/2$. (Il punto esclamativo indica il fattoriale.) Quindi per quattro monete vi sono tre permutazioni, per cinque ve ne sono 12, per sei ve ne sono 60 e così via.

Golomb propone un interessante problema irrisolto. Date n monete di differenti grandezze, quali procedure minimizzeranno e quali massimizzeranno le dimensioni della circonferenza racchiusa al loro interno? Golomb dispone di alcune congetture su entrambi gli algoritmi che presto compariranno in un articolo sull'argomento.

Un problema completamente differente e piuttosto poco noto, sempre a proposito di circonferenze tangenti, riguarda la disposizione di n circonferenze identiche in modo che non si sovrappongano all'interno di una superficie di una data forma: si vuole che l'area di quest'ultima sia la minima possibile. Il problema ha degli aspetti pratici dato che si presenta spesso la necessità di imballare bottiglie o barattoli in involucri a sezione circolare, quadrata o di altro tipo. Qual è l'area minima della sezione trasversale che permette di imballare n cilindri? Formulando il problema in un altro modo, data l'area di una regione e n circonferenze uguali, qual è il diametro massimo delle circonferenze che permette la loro disposizione nella data regione senza sovrapposizioni?

Non si conosce una soluzione generale, anche quando la forma della regione è semplice, come una circonferenza, un quadrato o un triangolo equilatero, e in ogni caso le disposizioni ottimali si conoscono solo per valori molto bassi di n.

L'arbelo, o «coltello da calzolaio», di Archimede.

Quando la figura è una circonferenza, si conoscono le dimostrazioni solo dei casi in cui n varia da uno a 10; esse sono dovute a Udo Pirl e risalgono al 1969. I casi con n che varia da 2 a 10 (presi dall'articolo *Packing Cylinders into Cylindrical Containers* di Sydney Kravitz in «Mathematics Magazine», vol. 40, n. 2, marzo, 1967) sono mostrati nell'illustrazione di pagina 125. Il diametro minimo della circonferenza esterna è dato sotto ogni figura, supponendo che le circonferenze piccole siano di diametro unitario.

Kravitz fornisce le migliori tra le soluzioni che ha potuto trovare empiricamente per n che varia da 11 a 19. Il caso di $n=12$ è particolarmente interessante. Si potrebbe pensare che la disposizione serrata mostrata a sinistra nella figura in alto di pagina 126 sia la più densa, ma Kravitz ha trovato che la disposizione a destra è leggermente migliore. Michael Goldberg nell'articolo *Packing of 14, 16, 17 e 20 Circles in a Circle* («Mathematics Magazine», vol. 44, n. 3, maggio 1971) fornisce delle soluzioni migliori per i casi citati nel

La successione dell'arbelo.

titolo. La sua disposizione nel caso di $n=17$ è stata ulteriormente migliorata da George E. Reis («Mathematical Magazine», vol. 48, n. 1, gennaio, 1975) che presenta anche congetture per i casi da $n=21$ a $n=25$.

Un rompicapo diffuso in Francia alla fine dell'ottocento, «The Packer's Secret», consisteva di una scatola circolare contenente 12 gettoni che dovevano essere disposti all'interno della scatola in modo da formare una struttura sufficientemente rigida per non cadere quando la scatola venisse capovolta. L'illustrazione in basso a pagina 126 è una circonferenza della dimensione esatta per riprodurre il rompicapo con 12 penny degli Stati Uniti. Siete in grado di disporre queste dodici monete in modo che formino uno schema in cui nessuna moneta ha l'agio di spostarsi? La risposta sarà data il mese prossimo.

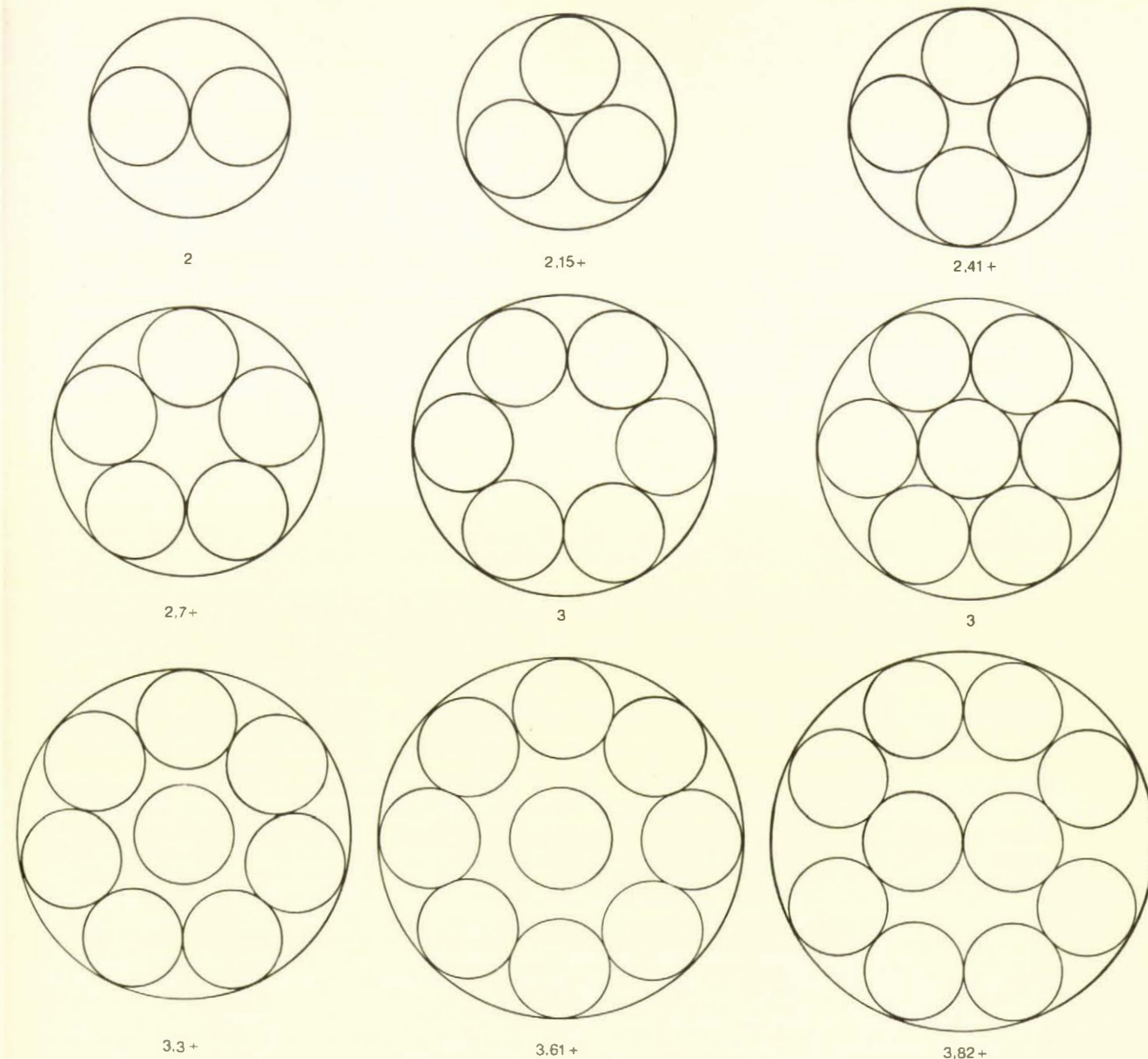
La catena di Jacob Steiner.

Una dimostrazione fallace per il problema di Solomon W. Golomb.

Si sono fatte ricerche anche per la disposizione più fitta di n circonferenze identiche entro quadrati. Si è dimostrato che col crescere di n la densità si avvicina a 0,9069+. Questo è il limite che si ottiene con la nota disposizione in cui i centri delle circonferenze formano un reticolo regolare di triangoli equilateri. Le dimostrazioni per le disposizioni migliori sono note solo per i casi di n che varia tra uno e 9. Anche qui non si conosce una formula o un algoritmo che fornisca la disposizione più fitta. L'illustrazione in alto a pagina 129, riprodotta da l'articolo *The Packing of Equal Circles in a Square* di Goldberg («Mathematics Magazine», vol. 43, n. 1, gennaio, 1970) mostra le soluzioni per n che varia tra 1 e 9. Nell'illustrazione, invece di collocare circonferenze unitarie entro il quadrato minimo, Goldberg mostra le circonferenze massime che si possono collocare in un quadrato unitario. Il suo articolo fornisce anche le congetture migliori per n che varia da 10 a 27 e per alcuni altri valori maggiori.

Goldberg dimostra che in ogni soluzione ottimale deve esistere una struttura di circonferenze tangenti che connette i quattro lati del quadrato e che entro questa struttura ogni circonferenza deve toccare almeno altre tre circonferenze oppure un lato del quadrato. Possono esistere tuttavia circonferenze che non fanno parte di questa struttura, come si vede nel caso di $n=7$.

Sotto a ogni quadrato nell'illustrazione è riportato il diametro massimo delle circonferenze nell'ipotesi che il lato del quadrato sia unitario. Nei due casi $n=6$ e $n=7$ le dimostrazioni non sono facili e non sono ancora state pubblicate. Il caso di $n=6$ è stato risolto per primo da Ronald L. Graham dei Bell Laboratories. Il caso di $n=7$ è stato provato a J. Schaer nel 1965. In una nota di Schaer pubblicata sul «Mathematics Magazine» (vol. 44, n. 3, maggio, 1971) è stata migliorata la congettura di Goldberg per $n=10$. Quando si tratti invece di disporre sfere entro sfere o cubi il problema diventa enormemente più difficile e su questo suo aspetto tridimensionale si conosce pochissimo. (A questo proposito si veda l'articolo di Goldberg in «Mathematics Magazine», vol. 44, n. 4, settembre, 1971.)

Anche la disposizione di n circonferenze uguali entro triangoli equilateri presenta problemi difficili. Si è ottenuto qualche risultato quando n è un numero triangolare (nella successione 1, 3, 6, 10, 15,...); in tal caso la disposizione più fitta si ottiene disponendo le circonferenze in righe di 1, 2, 3, 4, 5,....Se il numero di circonferenze è pari a un numero triangolare meno due o più unità, le circonferenze restanti si possono sempre spostare

Le disposizioni più fitte di circonferenze in una circonferenza: casi con n *che varia da 2 a 10.*

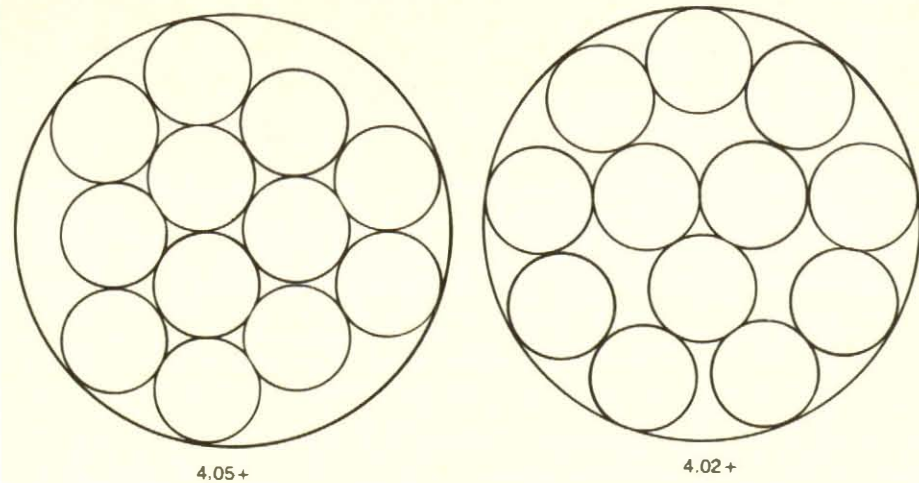
disponendole in un triangolo equilatero di dimensioni ridotte. Quindi se si tolgono due palle da un contenitore triangolare che ne imballa esattamente 15, le restanti 13 palle possono essere disposte in modo che nessuna tocchi il contenitore.

Supponete di toglierne solo una. Le restanti 14 possono essere disposte in modo che nessuna tocchi il contenitore? Sembra improbabile, ma non se ne conosce la dimostrazione. Donald J. Newman ha formulato una congettura secondo la quale in tutti i casi in cui n è uguale a un numero triangolare diminuito di una unità, non è possibile ridisporre le sfere (o le circonferenze) in modo da racchiuderle in un triangolo più piccolo. La congettura non si applica al numero triangolare 1 e vale ovviamente per 3. Sembra che valga per 6, ma non conosco alcuna dimostrazione del fatto che cinque palle non possono essere imballate in un triangolo più piccolo di quello che si usa per sei.

Un problema analogo si può porre a proposito della disposizione più fitta possibile di circonferenze uguali all'interno di una data regione, incluse le regioni con buchi. Se la regione non ha buchi ed è limitata da una curva chiusa convessa, il miglior risultato ottenuto finora è di Folkman e Graham in *A Packing Inequality for Compact Convex Subsets of the Plane.* («Canadian Mathematical Bulletin», vol. 12, n. 6, 1969): dati area e perimetro di una regione gli autori stabiliscono un limite massimo al numero di circonferenze unitarie che possono esservi disposte.

Dalle centinaia di teoremi riguardanti circonferenze tangenti voglio ricordarne un altro: un elegante risultato pubblicato nel 1968 dal geometra canadese H. S. M. Coxeter e illustrato nella figura in basso della pagina seguente. Si costruisce una successione infinita di circonferenze in modo che quattro circonferenze consecutive siano sempre tangenti. Si dimostra che questa successione è unica. Il raggio di ogni circonferenza si ottiene moltiplicando il raggio di quella più piccola più vicina per la somma del rapporto aureo con la sua radice quadrata, un numero che è leggermente maggiore di 2,89. I punti di contatto della circonferenza giacciono su una spirale equiangola tratteggiata.

Ecco le risposte alle domande del mese scorso:

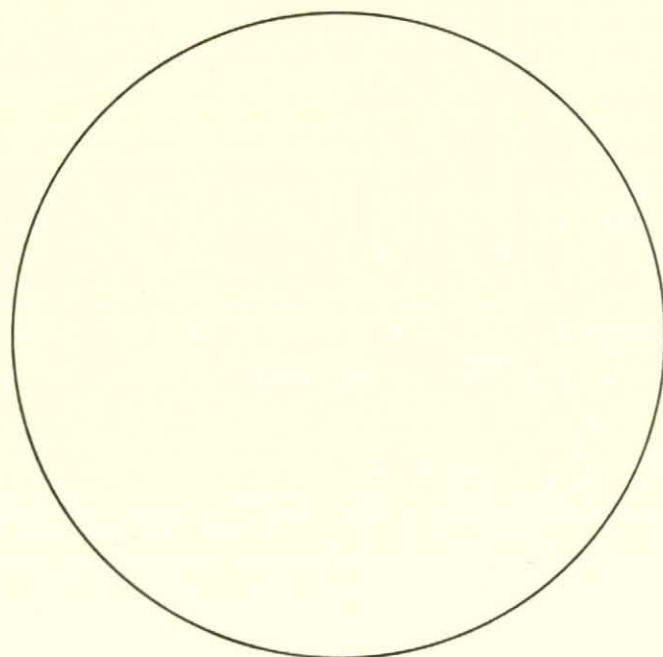
1. Inkstand.
2. Crankshaft.
3. Se gli interi positivi si dividono in pari e dispari, presa comunque una coppia in uno dei due insiemi essa darà come somma un numero pari maggiore di 2 e

Una disposizione serrata di 12 circonferenze (sinistra) *e quella che si ritiene la disposizione migliore.* (destra).

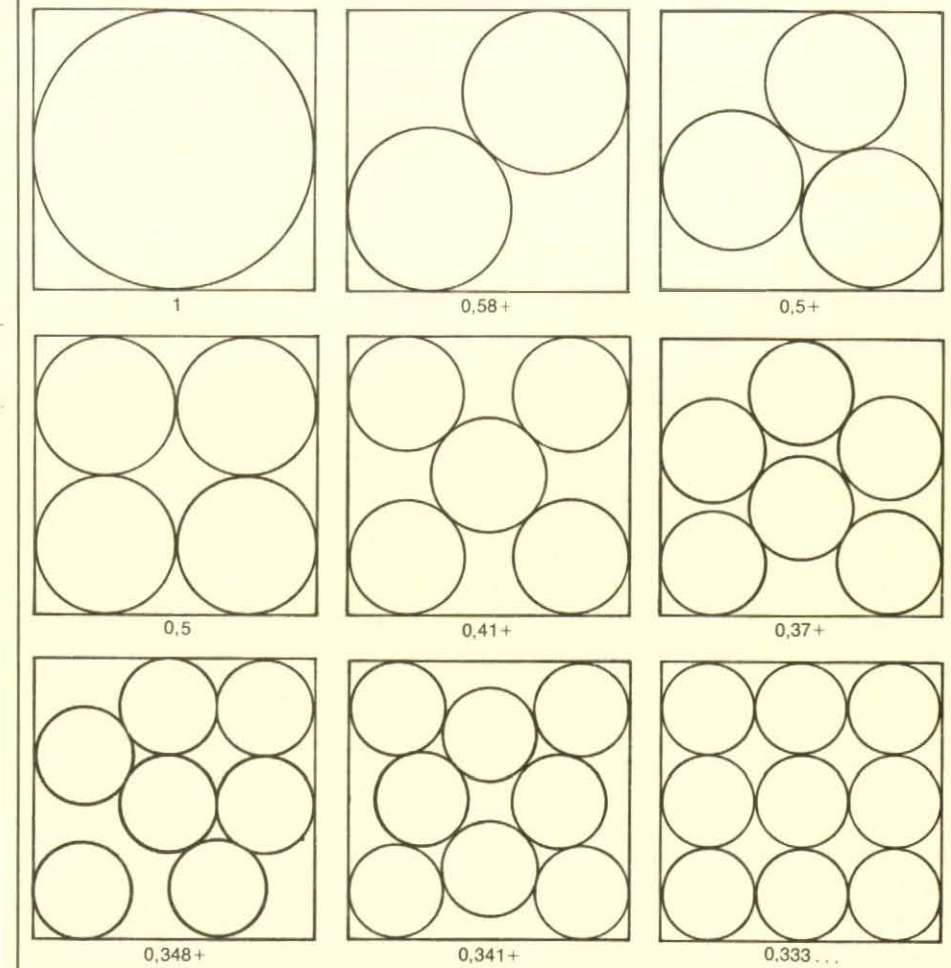
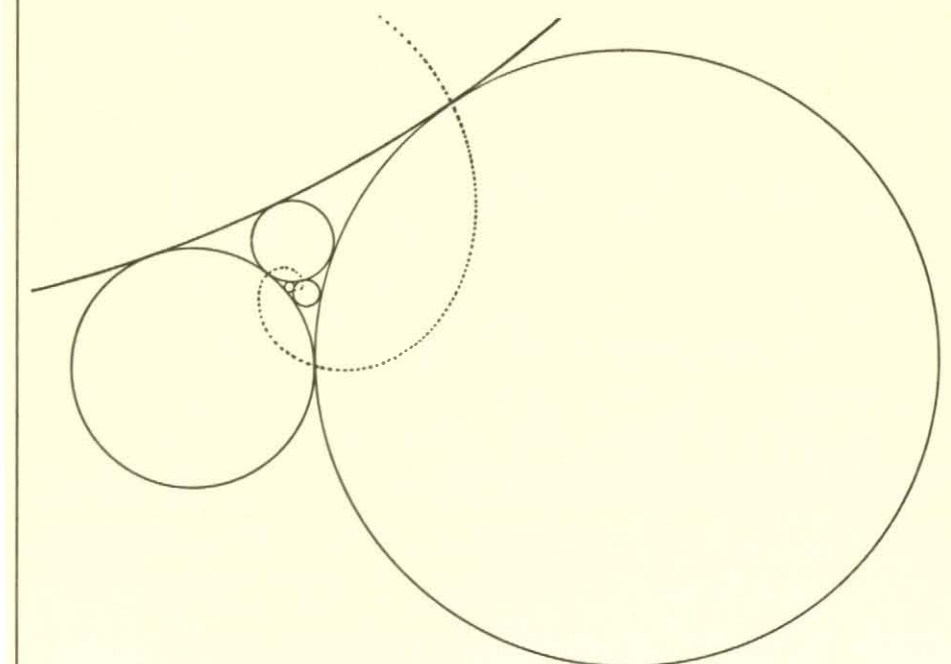
quindi non può dare come somma un primo. (Devo questo problema al defunto David L. Silverman.)

4. Il numero chiesto è 381-65-4729. Uno 0 posto alla fine dà l'unica soluzione del problema con le 10 cifre.

5. Una soluzione in 16 mosse al problema dello spostamento dei cavalli è la seguente. Come nella figura del mese scorso le righe sono numerate da 1 a 4 dal basso all'alto e le colonne sono indicate con *A*, *B*, e *C* da sinistra a destra.

1. *A*1-*C*2	9. *A*2-*C*1
2. *C*2-*A*3	10. *B*1-*C*3
3. *B*4-*C*2	11. *C*3-*A*4
4. *C*2-*A*1	12. *A*3-*C*2
5. *C*1-*A*2	13. *C*4-*A*3
6. *A*2-*B*4	14. *A*3-*B*1
7. *A*4-*C*3	15. *C*2-*A*3
8. *C*3-*A*2	16. *A*3-*C*4

Nell'articolo di febbraio sulle frazioni egiziane ho dato lo sviluppo in tre termini 5/121 = 1/25 + 1/759 + 1/208 725. Uno sviluppo in tre termini di questa frazione è il più breve possibile, ma mi ero chiesto se non fosse possibile darne uno con un minore denominatore massimo Sin Hitotumatu, Alan F. Berndt e Pierre Tougne sono stati i primi tra una lunga serie di lettori che hanno trovato lo sviluppo migliore: 5/121 = 1/33 + 1/121 + 1/363. C. S. Metchette, Morton Raff e R. Bruce Welmers sono stati i primi tra molti lettori a trovare il migliore degli sviluppi in quattro termini di 8/11. Ossia 8/11 = 1/2 + 1/6 + 1/22 + 1/66, che è una soluzione molto migliore di quella data alla fine del mio articolo di marzo. Nella quarta soluzione riportata in quello stesso articolo, *b* è uguale a 25 e non a 5. Entrambi gli sviluppi in quella soluzione sono per 3/25.

Circonferenza per giocare a «The Packer's Secret» con 12 penny degli Stati Uniti.

Sono felice di annunciare che un problema per lungo tempo rimasto irrisolto riguardante il ticktacktoe su una scacchiera tridimensionale di 4×4×4 è finalmente stato risolto. Con l'aiuto di un calcolatore, Oren Patashnik dei Bell Laboratories ha dimostrato ciò che molti avevano fino a quel momento semplicemente supposto, che il primo a muovere ha una strategia vincente.

Le disposizioni più fitte di circonferenze uguali entro quadrati con n che varia da 1 a 9.

La successione aurea di circonferenze tangenti di H. S. M. Coxeter.